University of Oxford

Genome Evolution and Epidemiology of Human Pathogens

Bethany L. Dearlove
St. Cross College

Nuffield Department of Medicine

A thesis submitted for the degree of Doctor of Philosophy

October 2013

Genome Evolution and Epidemiology of Human Pathogens

Bethany L. Dearlove, St. Cross College
D.Phil. thesis, Michaelmas Term 2013

Abstract

Understanding the transmission dynamics of infectious diseases is important to well-informed public health policy, responsive infection control and individual patient management. The on-going revolution in whole-genome sequencing provides unprecedented resolution for detecting evidence of recent transmission and characterising population-level transmission dynamics. In this thesis, I develop and apply evolutionary approaches to investigating transmission, focusing on three globally important pathogens.

Hepatitis C virus (HCV) is a major cause of liver disease affecting 150 million people and killing 350,000 annually. I conducted a meta-analysis of twentieth-century HCV epidemics, finding that the age of the epidemic can be predicted by genetic diversity. Using the coalescent, I fitted classic susceptible-infected (SI), susceptible-infected-susceptible (SIS) and susceptible-infected-recovered (SIR) epidemiological models. Most epidemics showed signatures of SI dynamics, but three, from Argentina, Hong Kong and Thailand, revealed complex SIR dynamics.

Norovirus is the leading viral cause of diarrhoea, estimated to cost the NHS around £115 million annually. I analysed whole norovirus genomes via a stochastic transmission model, finding that up to 86% of hospital infection was attributable to transmission from another patient in the hospital. In contrast, the rate of new introductions to hospital by infected patients was extremely low (<0.0001%), underlining the importance of ward management during outbreaks.

Campylobacter is the most commonly identified cause of bacterial gastroenteritis worldwide. I developed a zoonotic transmission model based on phylogeography approaches to test whether three strains previously associated with multiple host species were in fact aggregates of strongly host-restricted sub-strains, or genuine generalists. Members of the same strain isolated from different host species were often more closely related than those isolated from the same host species. I estimated 419, 389 and 31 zoonotic transmissions in ST-21, ST-45 and ST-828 respectively, strongly supporting the hypothesis that these strains are adapted to a generalist lifestyle.

In memory of my Grandma, with whom I shared a passion for reading and knowledge. She loved to talk about my studies and life in Oxford, but sadly passed away after a brave battle with leukaemia shortly before this chapter of my life was complete.

Pamela Jenner
24.03.1938 – 11.07.2013

Acknowledgements

Thanks go to the many members of the Modernising Medical Microbiology Consortium, who have been willing to share their vast expertise and experience. In particular, I would like to thank: Nicholas Wong, Jessica Hedge, David Wyllie, and Tim Peto for many useful discussions about norovirus; Richard Everitt for his help with Bayesian methods; and Liz Batty, Tanya Golubchik and Camilla Ip for their guidance in bioinformatics. I would also like to thank Madeleine Cule for her help with the norovirus transmission model, Samuel Sheppard and Guillaume Méric at the University of Swansea for inviting me to join their work on *Campylobacter* and allowing me to lead the subsequent analysis, and Karen Ayres at the University of Reading for inspiring me to enter this field. I am grateful to my family and friends who have supported me and helped put things into context throughout my D.Phil. journey: particularly Kate and Ian Dearlove, and Douglas Whalin, who have all been there with love and cups of tea to keep me going; and to the members of OUWPC who gave me a social life away from the computer. Finally I would like to thank Daniel Wilson for his guidance, insights and inspiration.

This work contained in this thesis was funded by the Nuffield Department of Medicine, University of Oxford and the National Institutes of Health and the Modernising Medical Microbiology Consortium under the UK Clinical Research Collaboration Translational Infection Research Initiative supported by the Medical Research Council, Biotechnology and Biological Sciences Research Council, National Institute for Health Research and the Wellcome Trust. I would also like to thank St. Cross College for financial support in attending SMBE2013 in Chicago.

Thanks also to my examiners, Mark Pallen and Chris Spencer.

Declarations

This thesis is entirely my own work, undertaken with the guidance of my supervisor Dr Daniel Wilson, except as detailed below.

Chapter 3

Mathematical formulation of the combined epidemiological-coalescent approach in this chapter was derived by Dr Daniel Wilson. I designed and conducted the literature review, collated and cleaned published data, performed data analysis and interpretation of hepatitis C epidemics, and wrote the thesis chapter. This work is published in Dearlove and Wilson (2013).

Chapter 4

The norovirus transmission model used in this chapter was designed and implemented in Python by Dr Madeleine Cule (Cule and Donnelly, submitted). The norovirus whole genome sequences were collected and prepared for RNA sequencing by Dr Nicholas Wong. The sequences were assembled using bioinformatics software by Dr David Wyllie (Wong et al. 2013). Anonymised patient data was supplied by Prof Tim Peto. I aligned the sequences, prepared the data for analysis, ran the analyses, interpreted the results and wrote the thesis chapter.

Chapter 5

The whole genome sequence alignments of *Campylobacter* were provided by my collaborators Dr Samuel Sheppard and Dr Guillaume Méric of the University of Swansea. I cleaned the data, devised and implemented the analysis, interpreted the results and wrote the thesis chapter.

Word count: 46,071

Publications

Dearlove B, Wilson DJ (2013) Coalescent inference for infectious disease: meta-analysis of hepatitis C. Philos Trans R Soc B 368:20120314. doi: 10.1098/rstb.2012.0314

Wong THN, Dearlove BL, Hedge J, et al. (2013) Whole genome sequencing and de novo assembly identifies Sydney-like variant noroviruses and recombinants during the winter 2012/2013 outbreak in England. Virol J 10:335. doi: 10.1186/1743-422X-10-335

Conference Papers

Dearlove BL, Méric G, Wilson DJ, Sheppard SK (2013) Host switching in *Campylobacter*. SMBE 2013: Symposium on diversity and evolution of microbes and their genomes, Chicago (Poster).

Wong N, Dearlove BL (2012) Norovirus transmission: point source or multiple re-introduction? UKCRC Modernising Medical Microbiology Conference, Oxford.

Dearlove BL, Wilson DJ (2012) Coalescent inference for infectious disease: analysis of hepatitis C. SMBE 2012: Symposium on microbial genome evolution, Dublin (Poster).

Dearlove BL, Wilson DJ (2012) Modelling the growth and transmission of infectious disease: linking epidemiology and population genetics. PopGroup45, University of Nottingham.

Glossary of Abbreviations

bp	Base pair
cDNA	Complementary DNA
CDS	Coding sequence
DNA	Deoxyribonucleic acid
HCC	Hepatocellular carcinoma
HCV	Hepatitis C virus
kb	Kilobase
Mb	Megabase
MCMC	Markov Chain Monte Carlo
ML	Maximum likelihood
μm	Micrometre
MLST	Multilocus sequence typing
MRCA	Most recent common ancestor
mRNA	Messenger RNA
NHS	National Health Service
nm	Nanometre
NS	Non-structural
OUH	Oxford University Hospitals
ORF	Open reading frame
PCR	Polymerase chain reaction
pdf	Probability density function
PFGE	Pulsed-field gel electrophoresis
PP	Posterior probability
RNA	Ribonucleic acid
RdRp	RNA-dependent RNA polymerase
SI	Susceptible-infected
SIR	Susceptible-infected-recovered (removed)
SIS	Susceptible-infected-susceptible
ST	Sequence type
tRNA	Transporter RNA
UTR	Untranslated region

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Motivation

The title of this thesis is 'Genome Evolution and Epidemiology of Human Pathogens' - bringing together the fields of infectious disease, epidemiology, genomics and evolution. In the past, these have been studied as separate fields, but the vast amounts of pathogen sequence data now available have the ability to revolutionise how we understand infectious disease dynamics, and therefore they are rapidly converging. This is where the focus of this thesis lies: in developing combined epidemiological and population genetic methods, with the view to understanding the dynamics of infectious disease from the initial infection all the way to end of an epidemic. In this chapter, I first provide a brief overview of each subject in its own right, before outlining what this thesis aims to add.

1.2 The Nature and Importance of Infection

1.2.1 Brief History of Infectious Disease

Modern understanding of infectious disease is based on germ theory. This was developed in the nineteenth century, and radically altered man's relationship to public health (Gilchrist 1998; Mendelson 2002). The theory developed gradually, and a deeper discussion of its development and the main players is given by Gaynes (2011). Anthony van Leeuwenhoek's pioneering development of the microscope in the late seventeenth century allowed him to describe and investigate single-celled organisms, which he named animalcules, and which we now know as microorganisms. However, his

reluctance to publish restricted spread of this knowledge until decades after his death (Dobell 1932; Gaynes 2011). It was not until Agostino Bassi that a link was first proposed between microscopic cells and infectious disease in humans (Porter 1973); his germ theory soon found support from the experiments and observations of Louis Pasteur in the early 1860s, and Pasteur presented germ theory to the French Academy of Sciences on April 29th, 1878 (Pasteur et al. 1878). Robert Koch brought the theory of germ-spread disease to the attention and practice of the wider medical community, and is known in bacteriology for his four postulates – a series of scientific principles used to establish the aetiological relationship between a particular microorganism and a disease (Koch 1878; Koch 1891). From this point, germ theory has been a widely accepted explanation for human disease, and further developments, such as the revolution of surgery practices by Joseph Lister, have all occurred in the context of a scientific community which has come to accept this theory.

Infection has impacted upon global history in dramatic ways over the ages (Dobson and Carper 1996a). The outbreak of the bubonic plague (*Yersinia pestis*) between the sixth and eighth centuries, known as the Justinian Plague, contributed to the end of the Late Antique world, the rise of Islam, and ultimately contributed to the rise of the Carolingian state (Little 2007; Bos et al. 2012; Harbeck et al. 2013; Wagner et al. 2014). The Black Death, a subsequent *Y. pestis* pandemic, killed up to half of the European population and up to 200 million worldwide in the thirteenth century, and is seen historically as one of the definitive transitions between the medieval and renaissance worlds, ushering in new economic and social orders (Benedictow 2004). The 1918

influenza pandemic killed more people than died from all causes in World War I (1914-18), with estimates upwards of 50 million (Johnson and Mueller 2002; Taubenberger and Morens 2006). Today, despite improved scientific understanding and refined global capacity to combat outbreaks, infectious disease and its treatment remains an important factor in shaping human society. Whilst some diseases such as smallpox have been all but eradicated (Henderson 1976), others continue to be a major threat to public health: human immunodefiency virus (HIV) has infected 35 million people and has no known cure, a third of the global population is infected by tuberculosis with ~8 million new cases annually, and antimicrobial resistance is increasing at an alarming rate in bacterial pathogens (Bloom and Murray 1992; Neu 1992; Fauci 2001).

1.2.2 What Causes Infectious Disease?

There are five main groups of pathogens that can cause disease: fungi, protists, bacteria, viruses and prions. Of these, only fungi, protists and bacteria are considered living organisms in a conventional sense, able to autonomously replicate by copying their nucleic acid-based genetic code using their own cellular machinery. Viruses (Ivanowski 1892; Beijerinck 1898) have a nucleic acid-based genetic code commonly contained within protein and lipid layers, but they are obligate parasites that are unable to replicate without co-opting the machinery of a host cell. Prions (Prusiner 1982) do not have a genetic code at all, taking instead the form of a communicable misfolding of a host-encoded protein. By triggering other susceptible proteins in healthy cells to misfold they cause damage by the build-up of protein plaques, often leading to cell death

(Prusiner 1991). In this thesis, I will consider three bacterial and viral pathogens of global significance, so I will explain these groups in greater detail.

Bacteria are prokaryotes: single celled organisms that have a less complicated structure than the eukaryotic cells which make up animals, plants, fungi and protists. Bacteria are smaller than eukaryotes, lack organelles such as mitochrondia, and contain a circular nucleic acid-based genetic code not contained within a nucleus. Other features of bacteria include a cell wall for protection; plasmids, which carry genes that are not essential for reproduction such as antibiotic resistance; pili, which are extracellular filaments involved in the attachment of the cell to other bacteria or the host; and flagella, which are similar to pili and help with locomotion (Berg and Anderson 1973; Bullitt and Makowski 1995).

Many of the pathogenic bacteria were classified in the late nineteenth and early twentieth century according to readily observable traits (Schleifer 2009) including the colony structure (clusters, pairs, individual cells), cell shape (for example, rod shaped bacilli, spherical cocci, coiled spirochaetes), the structure of the cell wall (susceptible to staining using Gram's method (Gram 1884) or not), the environment in which the bacteria were found and their growth requirements (e.g. aerobic, anaerobic). These classifications were summarised for easy reference in Bergey's Manual of Systematic Bacteriology, and bacterial nomenclature formalised (Buchanan et al. 1947). It was soon apparent, however, that such methods did not sufficiently resolve closely related taxa, with problems often occurring due to features being lost and gained when culturing the

same strain, and the effects of lateral gene transfer (Ferguson Wood 1949). Accordingly, modern classification increasingly turned to molecular techniques such as the guanine-cytosine ratio, genome-genome hybridisation and sequencing of the 16S ribosomal ribonucleic acid (RNA) gene (Wayne et al. 1987; Woese 1987; Olsen et al. 1994; Cho and Tiedje 2001). In practice, modern classification relies on a combination of traditional techniques and genetic sequencing. With the advent of whole genome sequencing (see Section 1.5) unprecedented levels of information are available to aid taxonomy, and this is reviewed in Thompson et al. (2013).

Viruses are generally much smaller than bacteria. They are acellular, containing their genetic material within a protective protein capsule, and thus rely on host cells to be able to replicate. Classification of viruses was standardized by the Baltimore classification system (Baltimore 1971). This scheme uses a number of key features of the nucleic acid in the viral genome – deoxyribonucleic acid (DNA) versus ribonucleic acid (RNA), single versus double stranded, positive or negative sense – and the method of replication (Table 1.2.1). The sense of a virus genome refers to how it is read. Positive sense RNA genomes can be read directly by the ribosome, the cellular apparatus that translates the genetic code into protein sequences. In this respect it is similar to messenger RNA (mRNA) expressed by the host cell. Negative sense RNA genomes are complimentary to the messenger RNA, and need to be transcribed into positive sense RNA using their RNA-dependant RNA polymerase before they can be translated. Both viruses studied in this thesis, hepatitis C virus and norovirus, are negative sense RNA viruses, and replicate in this way. Retroviruses such as HIV (Group VI) first reverse

transcribe their positive-sense RNA into DNA, which is then spliced into the host genome for reproduction.

Table 1.2.1: The Baltimore classification of viruses (Baltimore 1971).

Group	Features
I	Double stranded DNA
II	Positive sense single stranded DNA
III	Double stranded RNA
IV	Positive sense single stranded RNA
V	Negative sense single stranded RNA
VI	Positive sense single stranded RNA, with replication through a DNA intermediate in the life cycle.
VII	Double stranded DNA, with replication through a DNA intermediate in the life cycle

1.2.3 Epidemiology

The term 'epidemiology' is derived from the Greek *epi,* upon and *demos,* the people, and thus translates as 'the study of what is on the people'. As a modern subject, it represents the study of the distribution of disease and health, and their determinants, in the human population (though it can equally be applied to other species). Taken a step further, this definition includes the prevention, surveillance and control of disease and also extends to environmental and genetic factors (Susser 1979).

The modern study of epidemiology can be dated to the mid-nineteenth century. During a cholera outbreak in London in 1854, John Snow pioneered a methodology of observation, isolation and intervention that was able to identify a water pump as the focus of infection, and which has contributed to medical efforts to combat the spread of

disease ever since (Snow 1855). William Farr, a physician and statistician who worked alongside Snow, studied mortality in a range of demographic groups and their relationship to disease in his position as Registrar General; the UK Office for National Statistics has developed directly from his work (Farr 1852; Langmuir 1976; Webb and Bain 2011). Florence Nightingale is notable for her role in observing, recording (in particular, her polar-area graphs) and responding to her findings during the Crimean war, creating a plan to reform nursing as a result (Nightingale 1858; Nightingale 1859; Nightingale 1860). Over the course of the late nineteenth and into the twentieth century, epidemiological practices became informed by broader developments in the scientific community; Darwin's theory of evolution (see Section 1.3.1), Pasteur's theories concerning microbes (Section 1.2.1), and advances in genetics, have all improved understanding of the biology of infectious disease. By definition, epidemiology requires the recording and interpretation of public health data, and thus, while it is considered a branch of science in its own right, it has a long history of fruitful cross-fertilization of ideas with the principles and methods of statistics (discussed in much greater detail than afforded here in Sections 2.2.1 and 2.3) informing the medical community's ability to describe and predict the behaviour of complex populations and groups. Now, thanks to the greater resolution of information that widely-affordable genetic sequencing has provided, in combination with descriptive mathematical models, we are at the cusp of a new age in society's ability to predict and react to infectious outbreaks (see Section 1.6).

1.2.4 Nucleic Acids, Amino Acids and Proteins

The structure of a nucleotide is made up of three main parts – a sugar, a phosphate and a base. There are four possible nucleotides at a given site in a single strand of DNA, each consisting of the sugar-phosphate backbone covalently-bonded to one of four bases – adenine (A), cytosine (C), guanine (G) and thymine (T). They pair A to T and C to G via hydrogen bonds to form a double-stranded helix (Watson and Crick 1953). In the case of (typically single-stranded) RNA, thymine is replaced by uracil (U). The nucleic acid carries the information necessary for life, providing the template to synthesising many important components an organism needs to reproduce (Benzer and Champe 1961; Crick et al. 1961). This template is read in groups of three bases known as triplets as mRNA, and these are recognised as code for a particular amino acid by the ribosomes in the cell. Amino acids are transported to the ribosome by transporter RNA (tRNA), and matched to the codon on the mRNA by complimentary base pairing (for example, CGG would match to GCC).The amino acids are then joined together to synthesise the protein (Yanofsky et al. 1964). The central dogma of molecular biology, promoted by Francis Crick in 1958, asserts that nucleic acid can be transcribed into proteins (either directly from RNA, or from DNA via RNA), but not vice versa (Figure 1.2.1) (Crick 1970).

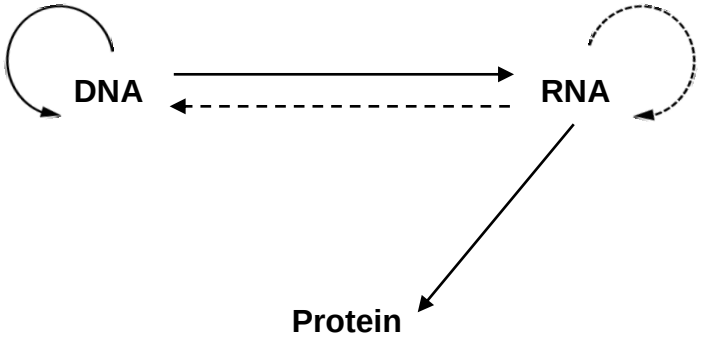


Figure 1.2.1: The central dogma of molecular biology. DNA (through RNA) and RNA can transcribe to protein (solid line), but protein cannot be translated back into nucleic acid. Dashed lines represent transfers that have only been observed under special conditions, for example, in the laboratory.

There are 20 possible amino acids, and three stop codons. The genetic code is therefore degenerate, as there are 64 possible triplets of nucleotides and thus more than one triplet can code for a single amino acid (Crick et al. 1961). A coding sequence (CDS) is a sequence of nucleotides that starts with a start codon (usually methionine, ATG), and finishes with a stop codon (TAA, TAG, TGA). Frequently in eukaryotes, and infrequently in prokaryotes, it may consist of multiple exons (coding regions) separated by introns (non-coding regions) and spliced together into a contiguous RNA transcript during gene expression (Forsdyke 2006). The CDS is usually identified through the mRNA (cDNA) sequence, and is an important step in annotating and understanding functional genes (Furuno et al. 2003). An open reading frame (ORF) is a bioinformatic prediction of a CDS from the DNA sequence by scanning for start and stop codons, but has not been confirmed as coding.

1.2.5 What is a genome?

The genome refers to the heritable genetic material of an organism, encoded by nucleic acid, including coding and non-coding regions (see Section 1.2.4). The unit of the genome is a chromosome (Morgan et al. 1915). Frequently in eukaryotes, and occasionally in prokaryotes, there are multiple chromosomes. Homologous chromosomes may be present in multiple copies, and the number (known as the *ploidy*) varies between taxonomic groups. Humans, for example, have 23 pairs of chromosomes in the nucleus, and thus have a ploidy of two (diploid), having received one copy of each chromosome from their mother, and one from their father (Tjio and Levan 1956). In contrast, bacteria and viruses are haploid, meaning that they only have a single version of the genome in their cells. The bacterial genome is usually circular and mainly contained in the nucleoid (free in the cytoplasm, unlike the membrane bound nucleus of eukaryotes), though some non-functional genes may be found in the cytoplasm on small circular chromosomes known as plasmids (Thanbichler and Shapiro 2006). Viral nucleic acid is more likely to be linear, and is sometimes surrounded by a protein capsid known as a nucleocapsid.

1.2.6 Molecular Epidemiology

Detailed investigation of infectious disease outbreaks relies on the identification of related cases. Traditionally, pathogens such as bacteria and viruses isolated from infected patients were characterized by observable traits, known as phenotypes – for example by size, shape and antimicrobial resistance profiling – and compared to determine close similarity or otherwise. With the advent of molecular typing methods,

such as gel electrophoresis, the field of molecular epidemiology was born. Molecular typing methods, particularly modern genetic typing methods including multilocus sequence typing (MLST) and whole genome sequencing (WGS), provide a rich source of traits on which pathogens can be classified and compared, allowing for better understanding of the relationships between isolates in investigating outbreaks and evolution (Gould 1986; Maiden et al. 1998; Medini et al. 2008).

Gel electrophoresis uses differences in the size and charge of molecules isolated from an organism to obtain a molecular signature that can be used in its classification (Smithies 1955; Thorne 1966). Originally applied to differentiate allelomorphs of the same enzyme (known as allozymes), nowadays pulsed field gel electrophoresis (PFGE) is used in conjunction with restriction endonucleases to digest DNA into fragments (Schwartz and Cantor 1984). The DNA fragments are then separated by passing an electric current through a gel in which the samples are suspended, during which the voltage of the gel is switched in three directions. This allows for the separation of larger DNA fragments, up to 2000 kilobases (kb) in length which can be used as a barcode for identification. However, there have been many studies citing difficulties in reproducibility and therefore comparison of results (Cookson et al. 1996; van Belkum et al. 1998). An overview of PFGE is given in Goering (2010).

From the late 1990s, MLST, based on dideoxynucleotide sequencing (see Section 1.5), provided greater sensitivity and reproducibility to detect bacterial diversity than had been achieved before, providing several advantages over other techniques including

serotyping, PFGE and flagellin typing (Harrington et al. 1997; Wassenaar et al. 1998; Steinbrueckner et al. 2001). MLST was originally demonstrated in typing *Neisseria meningitidis*, and distinguishes isolates based on 400-600 base regions of six to ten housekeeping genes (Maiden et al. 1998). MLST allowed insights in to more slowly accumulating global diversity, and was easily reproducible and shared between laboratories across the globe helping to ensure consistency; see Maiden (2006) for an in-depth review of MLST.

1.2.7 Genomic Epidemiology

In the last few years, inexpensive whole genome sequencing has begun to revolutionize molecular epidemiology, and in particular clinical microbiology (Pallen et al. 2010; Didelot et al. 2012a; Loman et al. 2012; Didelot 2013). Clinical microbiology has two main roles: diagnosis of disease, and the detection and monitoring of outbreaks (Didelot et al. 2012a). Despite the widespread use of molecular techniques in epidemiology, they remain time-consuming compared to the simple phenotypic tests that are available (Table 1.2.2), as they depend on the usual culturing time in addition to sequencing. For this reason, and their relative expense, they are not generally the first port of call in diagnostic microbiology laboratories, where taxonomic identification and antibiotic resistance profiling is often urgently needed for the care of critically ill patients.

However, whole genome sequencing has already begun to transform clinical microbiology (Pallen et al. 2010; Didelot et al. 2012a; Loman et al. 2012; Wilson 2012; Didelot 2013), offering an organism independent approach with rapid turnover and

ever decreasing costs (Section 1.5). Pilot studies have begun to illustrate just how real-time sequencing can change microbiology practice (Rasko et al. 2011; Rohde et al. 2011; Eyre et al. 2012; Köser et al. 2012). However, issues remain before it can be rolled out as a nationwide program (Robinson et al. 2013).

Table 1.2.2: Phenotypic methods for clinical microbiology.

Purpose	Available Phenotypic Methods
Taxonomic identification	• Gram staining (Gram 1884). • Observation of growth in selective media. • Coagulase testing (Sperber and Tatini 1975). • Antibody tests. • Matrix-assisted laser desorption/ionization–time of flight (MALDI–TOF) mass spectronomy (Seng et al. 2009).
Resistance typing	• Measuring growth in the presence of specific antimicrobials. • *MecA* testing in *Staphylococcus aureus* (Bode et al. 2012). • Hain testing in tuberculosis (Barnard et al. 2008).
Detecting determinants of virulence	• Antibody tests. • Polymerase chain reaction.

1.3 Evolutionary Biology

1.3.1 Natural Selection and the Tree of Life

Underpinning the interpretation of molecular and genomic epidemiology is the theory of evolution. To the extent that germ theory revolutionised understanding of infectious disease, Darwin's theory of natural selection (Darwin 1859) fundamentally changed how we view the entire living world. Ernst Mayr (1982) distilled Darwin's thesis in The Origin of Species down to a series of facts and inference:

Fact 1 – all species have great potential fertility and if all individuals born produced offspring, then their population size would continue to grow exponentially.

Fact 2 – despite fluctuations, populations usually remain stable.

Fact 3 – there are limited natural resources, but in a stable environment, they remain relatively constant.

Inference 1 – since more offspring are produced than can be supported by natural resources, and the population size stays stable, then there must be a struggle for existence with only a small number of offspring surviving each generation.

Fact 4 – every population shows a wide amount of diversity.

Fact 5 – this variation is passed from one generation to the next.

Inference 2 – survival in the struggle for existence is not random, but partly depends on the traits they have inherited. The survival being shaped in this way forms the process of natural selection.

Inference 3 – over time, this process causes a gradual change in the population, and leads to the production of new species.

In this way, the large differences in traits between species could be explained by an accumulation of smaller changes throughout time, and that the diversity and abundance of life can be traced back to a common ancestor through a branching process. This brought to the fore the concept of an evolutionary tree, where extant organisms (the leaves on the tree) are not distinct independent beings, but rather have descended from a common ancestor (the trunk or root of the tree) through a slowly changing continuum. This analogy of Darwin's theory was popularised by Ernst Haeckel, who drew several versions of the evolutionary tree of life and was the first to use the term 'phylogeny' (see, for example, Haeckel (1866)).

Darwin was not the first to propose a hierarchical organization of the natural world in the form of a tree, but he was the first to propose a plausible mechanism for how species diversify over time. Lamarck had proposed a two part system for evolution based on two laws: 1) 'use' increasing a trait, and 'disuse' reducing a trait, and 2) the inheritance of acquired characteristics, which could therefore accumulate over a number of generations to cause a large shift in trait. The most often cited example is that of giraffe's necks; Lamarck's theory was that since food is high up, the giraffe's neck gets longer to reach it, and then this acquired length is passed on, thus leading to even longer necks in subsequent generations (Lamarck 1809). Lamarck's view of the tree of life did not branch, per say, instead being made up of a number of parallel lines along which species would evolve from simple to complex.

Morphological traits had been considered for many years before Darwin to understand relationships between species. Carl Linnaeus developed an approach to organising species together in a hierarchy (domains are divided into kingdoms, which in are turn divided in to phyla and species). As a result, Linnaeus had a profound influence on biological nomenclature, giving unique identifying names to many species that continue to be used until this day (Linnaeus 1758). Systematic methodology for evolutionary trees, showing the ancestral relationships between species and taxa was pioneered by Willi Hennig (Hennig 1950; Hennig 1966). Cladistics, as the methods based on his approach are now known, groups organisms

together according to traits (including morphology, physiology, behaviour and ecology) that have been inherited from a distant ancestor.

The advent of molecular typing gave more traits on which a phylogenetic tree could be built (see Section 1.4), taking comparisons from less than 100 traits to 1000s of possibilities (Zuckerkandl and Pauling 1965; Delsuc et al. 2005). In particular, molecular observations of extremely highly conserved ribosomal RNA sequences (across both eukaryotes and prokaryotes) by Carl Woese gave rise to the hypothesis of three domains of life on the tree of life: *Eukaryota*, *Bacteria*, and *Archea* (Woese and Fox 1977; Woese et al. 1990). Here, life refers to cellular organisms, and there has been much debate as to whether the tree of life should include viruses, and if so, where they fit (Moreira and López-García 2009). Whilst a formal discussion is out of the scope of this thesis, it is worth mentioning that the discovery of viruses such as *Mimivirus*, which infects amoebae but resembles a gram-negative coccus bacteria (La Scola et al. 2003), has blurred the distinction between viruses and bacteria, adding heat to the debate (Raoult and Forterre 2008; Moreira and López-García 2009; Boyer et al. 2010).

1.3.2 The Modern Synthesis

In the early twentieth century, there were two schools of thought in genetics: saltationism and the biometry school. Saltationism, led by William Bateson, was based on the idea that Mendelian genetics could not be compatible with the gradual evolution of natural selection proposed by Darwin (Bateson 1894). The biometry school, including

Francis Galton, Karl Pearson and Walter Weldon, believed the opposite, and that the discrete nature of Mendelian inheritance could not explain their evidence of continuous evolution (Welden 1895; Galton 1897; Pearson 1898). In 1918, R. A. Fisher showed that it was possible to derive the observations found by the biometry school using Mendelian principles (Fisher 1918). From here, Fisher, J. B. S. Haldane and Sewall Wright all showed that natural selection could act within the framework of Mendelian inheritance (Fisher 1930; Wright 1931; Haldane 1932). This step in understanding is referred to as Neo-Darwinism or the modern synthesis, and between them, Fisher, Haldane and Wright founded the new discipline of population genetics, taking a mathematical approach to evolution.

The modern synthesis inspired a new way of thinking about inheritance and evolution, quickly spreading into evolutionary biology through Julian Huxley's book on modern systematics, and spawning much new practical research as a result (Huxley 1942).

1.3.3 Forces Driving Evolution

There are five forces of micro-evolutionary change: mutation, selection, drift, migration and recombination.

Mutation. Mutation provides the raw material for evolution. A mutation refers to any heritable change to the genome due to, for example, copying error or environmental damage. There are many types of mutation event that can take place, but in this thesis point mutations are of most relevance, since in general they will only have a modest

effect, and it remains possible to identify regions of homology for alignment (Section 1.4.1). In addition, they are amenable probabilistic modelling.

A point mutation is a change from one nucleotide base to a different one, for example C to T. This change might be genic or inter-genic. If it falls in a genic region, the mutation can be synonymous (does not affect the sequence of the amino acid), non-synonymous (does change the amino acid sequence and thus the resulting protein; for example, a substitution from …AAG… to …AAA… still reads as lysine, whereas …AAG… to …AAC… would change the amino acid to asparagine), or nonsense (causes a premature stop codon). Whilst not directly having an impact on protein synthesis, inter-genic mutations may have an effect in a regulatory region such a transcription factor binding site.

An insertion occurs when one or more additional nucleotides are added between two sites in the genome, such as …AATG… becoming …AACGCTG… A deletion is when one or more nucleotides are lost from the genome, such as …TAGCGT… to …TGT… When the number of nucleotides affected by an insertion or deletion is not a multiple of three and occurs in a genic region, the reading frame is shifted. This frameshift has a high chance of causing non-synonymous and nonsense mutations, the latter of which leads to protein synthesis being stopped prematurely.

Selection. Mutations that are beneficial (for example, a mutation conferring resistance against the antibiotic being used as treatment) will tend to increase in frequency in the

population. This is known as positive selection. Equally, mutations that are harmful are likely to leave fewer offspring in the population, and over time these mutations will tend to be lost from the population (negative or purifying selection). Mutations that have no effect on whether they will be inherited in the next generation are known as neutral. In pathogens, one of the strongest sources of selective pressure is the host immune system.

Genetic drift. The stochastic component of evolutionary change from one generation to the next is distinct from the deterministic component driven by selection and migration, and this stochastic process is known as genetic drift (Haldane 1924). Drift can cause the loss of otherwise beneficial traits simply due to fluctuations in the population. Beneficial traits are particularly susceptible to loss when they are at low frequency, such as when they first arise. Haldane gives the probability of a beneficial trait being lost by drift as approximately equal to $1 - 2s$, where s is the selective advantage (Haldane 1927). On the other hand, inherited mildly deleterious traits might persist and sweep through the population despite selection due to drift. The stronger the selective advantage or disadvantage of a trait in the population, the less it is affected by drift. The fate of neutral mutations in the population is a result drift (Kimura 1955).

Migration. Migration is the movement of an individual from one population to another. Migration can increase genetic diversity through the introduction of new alleles. This can drive changes in gene frequency. Migration can also limit divergence between

populations that have split by homogenizing them (Wright 1931; and see review in Rhymer and Simberloff 1996).

Horizontal gene transfer. In contrast to migration, horizontal gene transfer is the movement of genetic material from one bacterial cell to another, not through inheritance from mother to daughter cell (Freeman 1951). This can happen in three ways:

Transformation – uptake of nucleic acid in the environment.

Conjugation – transfer of plasmids or transposons via specialised pili.

Transduction – transfer via infectious bacteriophage.

Once foreign genetic material has entered the cell, it can be integrated into the genome via the mechanism of recombination. If it is integrated into the same location in the genome, this is known as homologous, otherwise it is known as illegitimate. From an evolutionary point of view, recombination is often used as shorthand to refer to all types of horizontal gene transfer, as well as similar processes in eukaryotes and viruses. Horizontal gene transfer is the mechanism through which antibiotic resistance can rapidly spread through a bacterial population (Naik et al. 1994; Varga et al. 2012).

Homology versus homoplasy

Homology describes a characteristic (or nucleotide, in the context of sequencing) that is the same or similar in two organisms due to it being inherited from the same common ancestor. In contrast, homoplasy implies that a trait was not inherited from a common ancestor. The term homoplasy was first used by Lankester, to described convergent evolution, whereby two species have evolved the same trait independently (Lankester

1870). Nowadays, traits whose patterns of inheritance are incompatible with the tree owing to repeat mutation, back mutation, or HGT (i.e. reticulate evolution) are all considered homoplasies, as in Maynard Smith and Smith (1998)). The use of homoplasy to detect recombination is discussed in more detail in Section 2.7.

1.4 Molecular Phylogenetics

1.4.1 Alignment-Based Approaches

The use of molecular traits for phylogenetics, as opposed to the phenotypic traits used in cladistics (Hennig 1950; Hennig 1966), was originally met with resistance by leading systematists including Ernst Mayr, Theodosius Dobzhansky and George G. Simpson (see review in Suárez-Díaz and Anaya-Muñoz 2008). Indeed, a debate about whether morphological or molecular methods were more informative about different evolutionary problems persisted into the 1990s (Patterson et al. 1993). However, molecular methods give greater resolution than possible with phenotyping, and in organisms such as viruses that do not have a fossil record offer the only window into the past, and thus have taken over in phylogenetic understanding (Gould 1986).

As seen in Section 1.3.3, there are many forces that give rise to variation in the genome. Molecular phylogenetics exploits the signals left in the genome from these processes, allowing regions of similarity in the genome to be used to infer homology. Homologous regions are then carefully aligned to specify the nucleotide-level patterns of common descent. There are many difficulties in aligning sequences, particularly if the sequences are relatively divergent, have a history of recurrent insertion and deletion, and consist of

long repetitive regions. The loss of homology due to processes such as insertion or deletion, is represented as a gap, '-'. In this way, alignment is a base-by-base representation of the homology between sequences.

There are many computational tools available to aid with alignment. A good overview of the numerous alignment tools, and the algorithms underlying them, is given in Table 3.1 of Higgins and Lemey (2009). Table 2 of Edgar and Batzoglou (2006) gives a range of scenarios, and which aligners are most suitable. In this thesis, I used two alignment programs. MUSCLE (Edgar 2004) is a good all round alignment tool, suitable for large a number of sequences (Blackshields et al. 2006). I also utilised the Geneious alignment tool in Chapter 3, as this allowed sequences to be easily collated and organised into a personal database (Kearse et al. 2012).

1.4.2 The Concept of a Molecular Clock

Before long, phylogenetic systematicists began to use patterns of molecular similarity to reconstruct, not only the relative degrees of relatedness between individuals, but also the relative divergence times between splits in the phylogenetic tree. Zuckerlandl and Pauling (1965) reasoned that the more traits had diverged along the branch of a phylogenetic tree, the more amount of time was likely to have passed. If one assumes that these traits accumulated at a consistent rate, then it is possible to obtain relative divergence times. However, this idea was considered controversial, with early critics highlighting that external forces such as selection and population size are not constant

over time, and therefore there was no reason for the accumulation of mutations to be constant over time too (see, for example, the review of Dietrich and Skipper 2007).
In his neutral theory of molecular evolution, Kimura developed a model of evolution that would predict a molecular clock (Kimura 1983). Firstly, he assumed that adaptation is rare, and therefore that most differences between species at the molecular level are not a result of positive selection. He also assumed that selection is strong, so that deleterious mutations do not contribute to divergence (due to being removed), and any polymorphism is neutral. Under these assumptions, Kimura showed mathematically that the rate of divergence is dependent on the mutation rate, and independent of population size (Kimura 1968).

From its outset, the molecular clock, and particularly the neutral theory, were controversial (Dietrich and Skipper 2007). Early objections included the question of whether the clock measured calendar time or generations, and the relative importance of neutral evolution versus adaptation. Empirical evidence shows that the strict molecular clock is not always biologically appropriate, and that rates of evolution vary across species, as shown for RNA viruses in Jenkins et al. (2002) and discussed more generally by Kumar (2005). There are many potential sources of rate variation, including generation time, replication mechanisms and selection (Kimura 1986; Bromham and Penny 2003). Inevitably, selection is of importance to pathogen populations, not least because there may be selective pressure imposed by changing environments, the host immune system, host jumps associated with zoonosis, and interventions such as antimicrobials. Although the strict neutral theory allows for adaptation, it assumes it is

rare and has no influence on genetic variability within species. With the advent of whole genome sequencing, the hundreds, or even thousands, of sequences available within a species, and the inherent within-species level of molecular epidemiology, puts strain on the assumptions underlying the neutral theory. Despite this, the strict molecular clock is often favoured as the simplest possible model in the absence of evidence against clock-like evolution, justified by Occam's Razor. Tests for strict clock-like evolution are available in many phylogenetic software including PAML (Yang 2007), Tree Puzzle, (Schmidt et al. 2002) and BEAST (Drummond et al. 2006; Drummond and Rambaut 2007).

1.4.3 Phylogenetic Trees

The concept of a phylogenetic tree for representing evolutionary relationships has been introduced informally already with Darwin's theory of natural selection and the cladistics of Hennig and his peers. Indeed, the only illustration included in On the Origin of Species by Darwin was a representation of evolution among species in the recognisable form of what we now define as a phylogenetic tree (Darwin 1859). In this section, I more formally introduce the mathematical concept of phylogenetic trees and how they can be reconstructed.

A phylogenetic tree is an acyclic graph which represents the relationships between sequenced individuals. Edges of the graph represent evolutionary lineages and an internal vertex represents the common ancestor of the edges that it joins. External nodes (i.e. those with only one connecting edge) represent extant individuals for which

sequences are available. Sequences that are more closely related will have a shorter distance or path between them on the graph. A clade is a sub-tree representing an ancestor and all its descendants. The branching structure of the tree, setting aside the lengths of those branches, is known as the topology. Two topologies are considered different if the only way of constructing one from the other involves removing one or more branches and relocating them elsewhere.

An unrooted phylogenetic tree is an undirected graph, which does not imply the direction of ancestry along a branch, and simply shows the amount divergence of individuals from common ancestors. In contrast a rooted tree is directed, with a concept of time from root to tip. A rooted tree can be obtained from an unrooted tree in two ways – by adding an outgroup (a taxon that is close to the individuals in the sample, but more divergent than the samples are from one another), or by applying a molecular clock for mid-point rooting (Felsenstein 1981).

Phylogenetic trees can be reconstructed either through distance methods, or using a probabilistic model of evolution (Saitou and Nei 1987; Huelsenbeck and Ronquist 2001; Swofford 2003; Criscuolo et al. 2006; Drummond and Rambaut 2007; Guindon et al. 2010). Distance methods start from a pairwise matrix of distances, either enumerated directly or corrected using an implicit evolutionary model, for example the Jukes-Cantor correction (Jukes and Cantor 1969). The resulting matrix is then used to reproduce the phylogeny via a hierarchical clustering approach such as the unweighted pair group method with arithmetic means (UPGMA) or Neighbour-joining (NJ) (Sokal and

Michener 1958; Saitou and Nei 1987). These clustering methods output only a single possible tree. Since the pairwise distance matrix is a lower dimensional summary of the whole sequence data, and clusters are found using a simple algorithm, such methods are computationally inexpensive, even for a large number of sequences (Kuhner and Felsenstein 1994). However, since only the distance matrix is retained from the original data, some information is lost and these methods therefore do not make the most statistically efficient use of the data for reconstructing the phylogeny. Furthermore, the ad hoc clustering algorithms employed by distance-based methods are not typically based on a principled, probabilistic model of sequence change, and therefore the phylogenies they produce are difficult to interpret from both an evolutionary and statistical standpoint.

In contrast, full sequence-based methods such as parsimony, maximum likelihood and Bayesian approaches retain the original information for each site in the alignment. Rather than based on a simple algorithm, these all use criterion-based optimization approaches and thus have to overcome the computationally intense problem of searching for the optimal tree. Parsimony methods attempt to find the topology with the minimum amount of evolutionary change required to recover the observed sequences, whereas maximum likelihood and Bayesian methods use a probabilistic substitution model (Huelsenbeck and Ronquist 2001; Drummond and Rambaut 2007; Guindon et al. 2010). The number of possible trees grows rapidly with the number of sequences (Cavalli-Sforza and Edwards 1967), so that it is infeasible to look at all possible trees for datasets with much more than 10 sequences (known as an NP complete problem). There

are various strategies of identifying reasonable parts of the tree space, including nearest neighbour interchange, sub-tree pruning and regrafting, and tree-bisections and reconnection (Felsenstein 2004). However, there is no algorithm that guarantees finding the optimal topology for more than a handful of sequences. The maximum likelihood and Bayesian methods used to infer phylogenetic trees in this thesis are described in greater detail in Section 2.8.

1.4.4 Gene Trees versus Species Trees

Phylogenetic methods were predominantly developed to understand the relationship between individuals in different species. However, in epidemiology the relationships that are of main interest are almost exclusively within a species. In fact, the topology of the tree relating individuals from different species may not be the same as the tree between the species themselves. This mismatch is known as the incomplete lineage sorting problem (see, for example, Avise et al. 1983; Pamilo and Nei 1988; Takahata 1989; Doyle 1992; Maddison 1997). It arises when the most recent common ancestor of individuals in the same species is not more ancestral than the split of the two different species (Figure 1.4.1). Therefore, it is possible for individuals to have a most common ancestor that is more closely related to a member of a different species than of the same species. One prominent example of incomplete lineage sorting is the separation of humans from orang-utans and gorillas (see, for example, Chen and Li 2001).

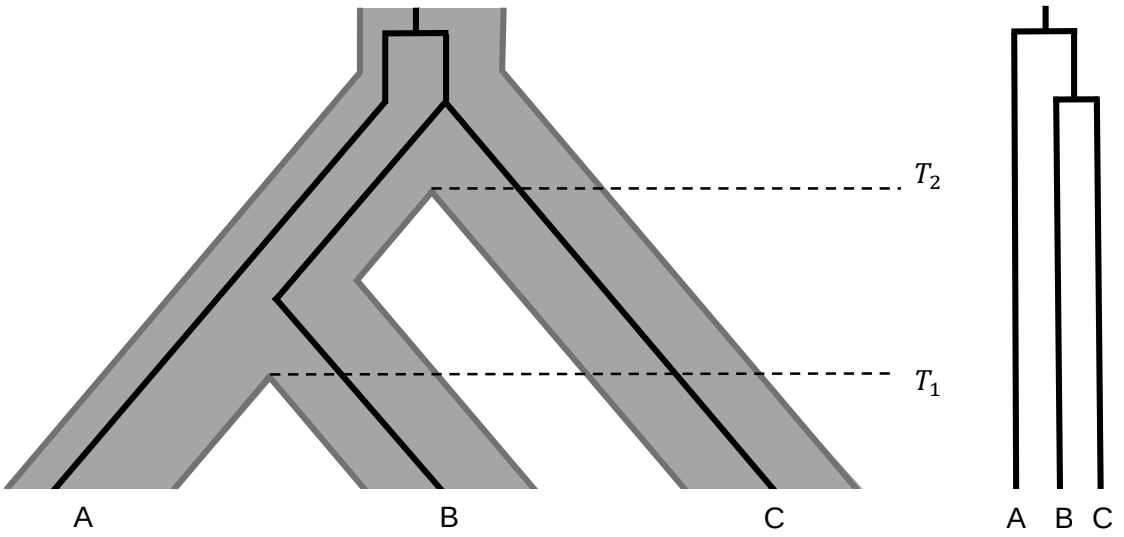


Figure 1.4.1: The incomplete lineage sorting problem. Given three species (A, B and C), the gene tree (black lines) can differ from the species tree (grey).

There are methods available that model the relationship within species, based on population genetics theory. In the next chapter I will go on to describe coalescent theory, an important branch of population genetics that allows mathematical models to be applied to drive evolutionary inference (Kingman 1982a; Kingman 1982b) and that I use in Chapters 3 and 5. Broadly, coalescent theory allows us to describe how processes at the population level affect the relatedness, or genealogical histories, underlying genetic samples and their variation, within a well-defined statistical framework, as a prior on a rooted tree. Coalescent inference is based on an explicit population genetics model, the standard neutral model. Importantly, this means that statements about biologically interpretable model parameters can be made, which represents an advance over post hoc interpretation of a phylogenetic tree. Although coalescent methods were developed to model relationships within species, they are sometimes applied at higher taxonomic levels (Degnan and Salter 2005; Liu 2008). An alternative phylogenetic method for

modelling the relatedness between different species is the Yule birth-death model (Yule 1924).

For more complicated evolutionary patterns, such as those that arise through recombination, the full evolution of samples cannot be represented simply on a tree. Instead, relationships need to be represented by a reticulated network, such as that modelled by the Ancestral Recombination Graph (ARG), an extension of the coalescent in the presence of recombination, or more generally a phylogenetic network (Hudson and Kaplan 1988; Griffiths and Marjoram 1996; Huson and Bryant 2006). These issues are considered in more detail in Chapter 5, Section 5.3.4.

1.5 Sequencing Methodologies

1.5.1 Sanger Sequencing

DNA sequencing originated in the late 1960s and 1970s, pioneered by Sanger (Sanger et al. 1965; Brownlee et al. 1967; Barrell and Sanger 1969; Sanger et al. 1973; Sanger et al. 1974; Sanger and Coulson 1975). In their seminal paper, Sanger et al. (1977) describe a method using termination nucleotides and electrophoresis that allowed sections of sequence to be read. The first step in sequencing protocol is to denature the DNA to make it single stranded; for RNA sequencing, a copy of complementary DNA (cDNA) is made from the mRNA. The polymerase chain reaction can be used to create more starting nucleic acid material if necessary. Single stranded DNA is then added to a reaction containing a termination nucleotide (say, for A), and all other bases as standard nucleotides (C, G and T in this example). The termination nucleotides differ from

standard nucleotides in that there is only a single hydrogen atom attached to the 3' carbon, as opposed to a hydroxyl group, OH. This means that new nucleotides can no longer attach to the chain, so the synthesis of the complementary strand is terminated as soon as a termination nucleotide is attached. Thus, at the end of this step, each reaction contains different length fragments all ending in the same base. These are then put into separate lanes, and separated by gel electrophoresis. Smaller fragments will move further through the gel, and the resulting bands on the plate can be read from smallest fragment to longest, with the nucleotide base at each site identified by the lane in which the band appears.

One of the limitations of Sanger sequencing as explained above is that the length of sequence able to be read is restricted by the distance from the starting end. Reads for sequences more than 1 kilobase (kb) in length tend to be cut short due to a terminal nucleotide being attached prematurely. Shot-gun sequencing overcomes this issue by first shearing or using a restriction endonuclease to break the DNA into smaller fragments (Anderson 1981). These fragments are then cloned into a vector, and a number of these templates are randomly sampled and sequenced according to the protocol described above. These fragments, known as reads, are then assembled, using the overlapping ends to help, similarly to the methods described in Section 1.5.3 (Staden 1979). Thus it was possible to use Sanger sequencing to look at whole genomes.

Sanger sequencing became the primary technology for sequencing, and a major achievement of the method came in 1995 with obtaining the first bacterial genome

Haemophilus influenzae, (Fleischmann et al. 1995). The development of radioactive and fluorescent markers for the terminating nucleotides allowed the possibility of reading bases optically (Smith et al. 1985; Smith et al. 1986). These markers mean that nowadays only one single reaction is needed, as opposed to one for each base; each nucleotide gives off a different wavelength of light, which can then be read automatically.

In the last decade there has been a rapid development of new, so-called second generation, sequencing techniques. These have allowed DNA to be sequenced at a much higher rate and lower cost, facilitating a wide range of new studies (Pallen et al. 2010; Loman et al. 2012). However, there are downsides, with these techniques prone to higher per base error rates and shorter read lengths (see, for example, Kircher et al. (2009) and Dohm et al. (2008)). Prominent technologies include the 454 Genome Sequencer (Roche Applied Science; Basel, Switzerland), the Illumina platform (Illumina; San Diego, USA) and the SOLiD platform (Applied Biosystems; Foster City, California, USA). Since Illumina sequencing was used to sequence isolates for this thesis, I will only explain that method in further detail.

1.5.2 Illumina High Throughput Sequencing

The development of the visual optics technology and processing was a key innovator in high throughput sequencing. The first step is library preparation, where the sample is purified and then the DNA is split into fragments using either restriction endonucleases, sonication, nebulization or shearing (Syed et al. 2009; Davey et al. 2011; Myllykangas et al. 2011). Adaptors are added to each end of these fragments, and then one end is bound

to a surface known as a flow cell. The free end then links to a matching adaptor also on the flow cell, forming a bridge. A bridge polymer chain reaction is used to amplify the DNA fragments (Adessi et al. 2000; Fedurco et al. 2006), with the resulting double stranded fragments being denatured to leave numerous clusters of similar single strands fixed to the flow cell at one end (Shendure and Ji 2008). To obtain the sequence from this single stranded DNA, a modified DNA polymerase and nucleotides are required. Each different type of nucleotide has a fluorescent label, and during each cycle of the sequencing reaction these are added to the sequence and photographed with a laser. A terminator ensures that only one base is read at a time, and is cleaved afterwards to allow the next base to be read in the sequence. This is repeated many times, until the entire cluster has been photographed (Metzker 2010).

As in shotgun sequencing, the resulting reads represent overlapping sections of the original DNA. To obtain the whole genome sequence, these reads need to be either mapped to a reference, or *de novo* assembled.

1.5.3 Read Mapping and Assembly

There are two main methods for obtaining a consensus genome from second-generation sequence reads: reference-based mapping, and *de novo* assembly. *De novo* assembly aims to reconstruct the sequence without any information on what it should resemble (in contrast to reference-based mapping, below), and thus of the two options is the more desirable. Software for this purpose includes Velvet (Zerbino and Birney 2008) and Cortex (Iqbal et al. 2012). *De novo* assembly can be compared to putting together a very

complicated jigsaw puzzle without a picture to refer to; difficulty arises not from lacking the picture, but due to many pieces being similar because of inherent repetitiveness in the genome. As a result, obtaining a whole genome is rare, instead often resulting in many isolated fragments of contiguous sequences, known as contigs, without any context of how they fit together. In bacterial genomics, where coverage per base is typically high, this is a fundamental problem; the presence of repeat regions (either tandem repeats, or multiple copies of the same gene in different parts of the genome) cause a large number of short contigs when reads are shorter than the length of the repeat. There may also be missing pieces owing to transient drop off in coverage, which itself varies depending on local factors including GC content. This can make comparing large numbers of sequences difficult, as they first need to be aligned (Section 1.4.1) and the fragments from different sequences may have been constructed differently and not overlap.

Mapping approaches exploit a high quality fully-sequenced, closed reference genome (often obtained by Sanger sequencing), to solve the problem that sequencing short reads alone may not allow the genome to be closed into a single contig. Reads are lined up against the reference using an algorithm, for example Stampy (Lunter and Goodson 2011) or BWA (Li and Durbin 2009). For each base in the alignment, the total number of reads is known as the coverage, and it may be the case that all, none, or some of the reads match the reference. A decision is then made as to which base is most likely to be correct and output in the final assembled genome using the base quality score. This is a

statistic concerning the error, calculated using information about the number of reads supporting that base, the sequencing quality, and the total coverage.

The main disadvantage to mapping is the reliance of the approach on the reference genome used. The best reference will be very similar to the sample; however this is not always possible, especially when assembling emerging outbreak strains or disparate strains. In addition, if the genome of interest has gained extra material relative to the reference genome, such as integrated mobile elements, then this information will be lost in the form of reads that cannot be mapped. A major strength of mapping comes when comparing many sequences, since they can be all aligned against the same reference, avoiding a large and computationally costly whole-genome multiple alignment step.

1.6 Using Genomics for Epidemiology

There are two main challenges in epidemiology that genomics can help study, and both are considered in this thesis. The first challenge is to resolve relatedness between samples isolated by different patients to help resolve transmission. Several proof of concept studies have begun to show how high throughput sequencing can be used for this purpose, but some thought is required to make best use of this data, and presenting it appropriately for use in the clinic (Rasko et al. 2011; Rohde et al. 2011; Eyre et al. 2012; Köser et al. 2012).

The second challenge is to use genetic information for estimation of epidemiological parameters to elucidate disease dynamics during the course of an outbreak. If

systematic monitoring has taken place throughout an epidemic, then important parameters such as the basic reproductive number (the average number of new infections caused by a single index case; denoted R_0), can be estimated directly, for example in the 2001 British foot and mouth outbreak (Ferguson et al. 2001), and in the severe respiratory syndrome outbreaks in Asia 2002-3 (Lipsitch et al. 2003). Such parameters have an important and time-dependant role in the understanding of unfolding disease dynamics, and informing potential intervention strategies (Anderson and May 1991). Alongside these traditional methods, genetic analysis previously has been used to gain insight into the epidemic history of a pathogen when reliable surveillance data is unavailable – as in the 2009 H1N1 influenza epidemic (Fraser et al. 2009; Smith et al. 2009) – or to add information about the origin of an outbreak in process, as in the Haitian cholera outbreak in 2010 (Chin et al. 2011). With real-time genomic sequencing becoming widely available in the near future, it is clear that genetics will have an increasingly important role in infectious disease outbreak understanding and response.

Genetic approaches are well-established for investigating the epidemic history of pathogen populations. Generally these methods stem from interpreting an evolutionary tree - where the topology may reveal unforeseen relationships between isolates and reveal transmission pathways (Cottam et al. 2008; Lieberman et al. 2011), and the shape may allude to overall population dynamics (Grenfell et al. 2004). More sophisticated models allow the genetic and epidemiological contributions to infectious disease to be explicitly modelled. One such example is the coalescent model, which is readily adapted

for different demographic models (see Section 2.5.3) and has been used to infer historical changes in population size (Slatkin 1977; Pybus et al. 2000; Nordborg 2008). In particular, these changes in population size have been used to infer historical prevalence in viruses (Pybus et al. 2001). More recently, whole genome sequencing has revealed new levels of diversity in bacteria, creating a wider need for combined population genetic and epidemiological inference (Wilson 2012).

1.7 Aims of the Thesis

This thesis presents three methods for joint genetic and epidemiological inference into the roles of transmission and evolution in human pathogens. In Chapter 3, I first explain a method for inferring population-level transmission rates based on integrated coalescent epidemiological models, before applying it to Hepatitis C epidemics from across the world to explore how genetic and demographic factors affect genetic diversity and transmission. This modelling is at a population level, and facilitates inference of the historical dynamics of an epidemic and formal comparison to other epidemics, for example to see the effect of different intervention strategies, or as here, to identify what is driving genetic diversity.

In contrast, Chapter 4 uses a stochastic approach to investigate transmission at the level of individual patients within a single NHS Trust in England. In this chapter I focus on a single strain of norovirus, and use a stochastic epidemiological model to reconstruct patient transmissions according to what is known about their ward movements and genetic information from whole genome sequenced samples. This reconstruction can

then be used to assess the relative importance of visitors, contamination and wards in the norovirus transmission chain, providing evidence towards useful strategies for reducing spread in future epidemics.

In Chapter 5 I consider transmission between populations of bacterial in different species, using a zoonotic model of transmission. I focus on a leading bacterial cause of gastroenteritis, *Campylobacter,* and exploit the power of whole genome sequencing to investigate zoonotic transmission in three strains that appear, on the basis of MLST, to be generalists capable of living in multiple hosts including chicken, cattle and pig. In addition, I infer the source population for a number of human isolates, providing whole genome support to the importance of the food production line.

Finally, in Chapter 6, I summarise the key results of this thesis. I will then highlight the uniting themes of this work, and put the study into the wider context by discussing developments in the field during the last three years and what the future holds.

Chapter 2: Methods for Analysis

2.1 Introduction

This thesis presents analyses for understanding the evolution and transmission of human pathogens. Whilst the individual techniques used within each chapter are diverse, much of the underlying theory is shared. This chapter reviews background theory and methods for the forthcoming results chapters. First, an overview of statistical inference is given (both maximum likelihood and Bayesian), and then the focus shifts to specific methods for understanding pathogen evolution and ancestral relationships, including software where appropriate.

2.2 Statistical Inference

In general, statistical inference involves the estimation of unknown parameters from a population of interest – for example the prevalence of infection at a certain time. Inference is made about these unknown parameters using data and a statistical model. There are two main schools of statistical inference, frequentist (or classical) and Bayesian (Fienberg 2006). A philosophical difference exists between the schools. In frequentist statistics, the unknown parameters are assumed to be fixed (Neyman 1937), whereas in Bayesian inference the unknown parameters are seen as random variables, so that uncertainty about their value can be modelled by some probability distribution (Bayes and Price 1763). While this distinction may appear slight, it leads to practical differences between frequentist methods (such as maximum likelihood) and Bayesian methods. Notably, Bayesian inference requires specification of a 'prior distribution'. This prior

distribution is chosen to reflect information (or lack thereof) that is known about the parameter before any data is observed, and can be based on previous modelling or expert opinion (O'Hagan 1998). Current knowledge about a parameter is updated using the observed data to obtain a posterior distribution representing the uncertainty surrounding parameters. An overview of the main statistical inference methods used in the rest of the thesis, both Bayesian and frequentist, is given in Chapter 2.

2.2.1 Maximum Likelihood

Maximum likelihood (ML) is a frequentist method for estimating parameters, by finding the values at which they give the highest probability of describing some observed data. The probability density function (PDF) describes the relative probability of observing any given value of the data, X, given the parameters θ, and is written by $f(X|\theta)$. To estimate the unknown parameters given an observed sample of data, the likelihood is first denoted as $L(\theta|X) = f(X|\theta)$. The likelihood is then maximised with respect to θ to estimate the unknown parameters. If the data are a sample of independent and identically distributed random variables $X = x_1, x_2, \dots, x_n$, then the maximum likelihood estimate of the parameters, denoted $\hat{\theta}$, is given by:

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} L(\theta|X) \\ &= \arg\max_{\theta} \left[\prod_{i=1}^{n} f(x_i|\theta) \right]. \end{aligned} \tag{2.1}$$

Often it is easier to find the maximum using the log-likelihood. This gives the same parameter estimate since $L(\theta|X)$ and $\ln L(\theta|X)$ are monotonically related and thus their maxima occur at the same value. For simple distributions, the maximum can be found

analytically, but when there are many parameters, this is not generally possible. Instead, numerical optimization methods are required to iteratively converge on a maximum.

2.3 Bayesian Inference

Underlying Bayesian statistics is Bayes' Theorem (Bayes and Price 1763). It states that, if the probabilities of events A and B occurring are greater than zero, then the probability of A conditional on B is given by

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}. \tag{2.2}$$

$p(A)$ can be thought of as the initial or prior probability of event A taking place, and $P(B|A)$ is the likelihood of B happening given A has already been observed. It is usually more conventional to denote the unknown parameters by the vector θ, and the data by X. As before, the likelihood function is equal to the probability of observing the data given the parameters. In addition, some prior information known about the parameters θ can be represented by the prior distribution, $p(\theta)$. As such, Equation (2.2) becomes:

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}. \tag{2.3}$$

The posterior represents a re-weighting of the prior information, $p(\theta)$, by the likelihood of the observed data, $p(X|\theta)$. It may be that very little prior information is available, and in this case it is usual to use a non-informative prior, such as a uniform distribution (all possibilities have equal probability) or a vague prior with a large variance to represent the wide uncertainty. The term on the denominator of Equation (2.3) is known as the marginal distribution of X. It is often hard to obtain, but since it is only a normalising constant it is usually sufficient to calculate:

$$p(\theta|X) \propto p(X|\theta)p(\theta) \tag{2.4}$$

i.e.

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

Since the posterior is a full distribution, it is necessary to summarize it in order to obtain point estimates and estimates of uncertainty. Possible point estimates include the mean, median and mode, which coincide for symmetrical unimodal distributions. A credible interval is useful to assess the range of distribution and accuracy of a point estimate. A $100(1-\alpha)\%$ credible interval is any interval $[a, b]$ such that the posterior probability between a and b is $100(1-\alpha)\%$. Many such intervals may satisfy this requirement. One possible interval is the range of values that lie between the $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles of the posterior distribution. In this thesis, 95% credible intervals will usually take this form, representing the 2.5% and 97.5% points. An alternative to using percentiles is the highest posterior density (HPD) interval, representing the narrowest interval $[a, b]$ containing $100(1-\alpha)\%$ of the posterior density.

The credible interval is sometimes used as a form of hypothesis testing. For example, if a credible interval does not contain zero, that may be considered appreciable evidence that the parameter is non-zero. Similarly, if two credible intervals are non-overlapping, that is usually considered to be good evidence that the parameters are different. It is worth noting here that credible intervals are different to confidence intervals in frequentist statistics. A $100(1-\alpha)\%$ confidence interval refers to the interval that, if calculated after repeated sampling, would contain the true value of the parameter

$100(1-\alpha)\%$ of the time. The Bayesian concept of a credible interval does not appeal to the concept of repeated sampling, but rather represents uncertainty in the parameter when considered as a random variable. Therefore it depends on the prior distribution.

2.3.1 Markov Chain Monte Carlo

In most cases, the posterior distribution cannot be solved analytically. Even if random samples can be drawn from the prior distribution, the parameter space in phylogenetics is, in general, extremely large due to the number of possible tree topologies. The chance of simulating a tree compatible with the observed data in this way is highly remote. Therefore the chance of sampling the posterior densely enough to gain useful inference is very small. Instead, approximate computational techniques can be used, such as Markov Chain Monte Carlo (MCMC). Informally, the idea is to sample a point from a known proposal distribution that is designed to produce parameters compatible with the observed data, and then decide whether this point is indeed sufficiently compatible with the target distribution.

A Markov chain is a stochastic process consisting of a set of random quantities defined in some state space θ, where the current state only relies on the state immediately prior to it (Markov 1971). This is known as the Markov property and, considering a draw of state $\theta^{(t)}$ at iteration t, implies that the previous state ($\theta^{(t-1)}$) and next one ($\theta^{(t+1)}$) are independent conditional on the current one, such that

$$p\left(\theta^{(t+1)}\middle|\theta^{(1)},\theta^{(2)},\dots,\theta^{(t)}\right)=p\left(\theta^{(t+1)}\middle|\theta^{(t)}\right). \tag{2.5}$$

Therefore Markov chains 'remember' only where they were are the last step. Under certain conditions, Markov chains can be constructed that will converge on some target distribution, and will do so regardless of where the chain started. The idea of MCMC is to devise a Markov chain whose stationary distribution will converge on the posterior distribution, and sample parameters from that distribution.

One method of constructing such a chain is via the Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970), and this is explained here as it is the algorithm that underlies the software BEAST used in later chapters. The chain is started at an arbitrary point $\theta^{(i)}$. The proposed value for the next state, denoted θ^*, is drawn from the proposal distribution, $q(\theta)$. The next step is to decide whether the current or proposed value for θ is most likely to have come from the posterior density. This ratio is used to calculate an acceptance probability for the proposed value:

$$p(\text{accept } \theta^*) = \min\left[1, \frac{p(\theta^*|X)q\left(\theta^{(i)}\middle|\theta^*\right)}{p(\theta^{(i)}|X)q\left(\theta^*\middle|\theta^{(i)}\right)}\right]. \quad (2.6)$$

If θ^* is more likely to be obtained under the posterior than the previous value $\theta^{(i)}$, the ratio will be greater than one and thus the probability of acceptance is taken to be one – meaning that a better proposal will always be accepted. Otherwise, the proposal is accepted with probability equal to the ratio of the target densities. If the proposal is rejected, then the chain stays at the current state. This strategy is repeated to obtain a long chain of values – the more iterations run, the more accurate the estimate of the posterior density will be. Informally, the algorithm works since more samples will be kept from high density regions of the posterior (due to the ratio in Equation (2.6)), and

thus the amount of time the chain spends in a particular parameter space once it has reached the stationary distribution is proportional to the posterior probability.

2.3.2 Metropolis Coupling

The aim of MCMC is to sample from across the whole parameter space; however if there are a number of disjoint regions, or modes, all with high posterior probability, the chain might get stuck in a local peak and not sample the other regions of similar density. This is known as a mixing issue, and often occurs when the parameter space is large and complex, such as that representing tree topologies. A useful method to detect mixing or convergence issues is to run two or more chains, and test whether the chains converge to different topologies, and this is done when using Bayesian methods in all three results chapters presented in this thesis.

Metropolis coupling (Geyer 1991) uses several parallel chains to improve mixing and therefore help make jumps between peaks, and is implemented in BEAST from version 1.7 onwards (Drummond et al. 2012). As an overview, the 'cold' chain samples from the posterior distribution, whilst 'heated' chains sample from a distribution of the posterior which has been 'flattened' by mathematically raising the distribution to a power less than one. These heated chains produce smoother distributions than the posterior, which facilitates mixing and swapping between peaks. Swaps are also proposed between two randomly chosen chains, and these are accepted in a manner similar to the Metropolis-Hasting acceptance probability. When a swap to the cold chain is accepted, this can allow jumps between local peaks, facilitating better mixing. Since only the cold chain

represents the posterior distribution, the samples from the heated chains are discarded at the end of the analysis and only samples from the cold chain are retained.

2.3.3 Diagnosing the MCMC

The MCMC algorithm needs to explore the whole parameter space efficiently, and it is important to ensure that this has been the case and that it has successfully converged to the stationary distribution. Tracer (Rambaut and Drummond 2009) is a program that helps to visualise the results from multiple MCMC chains, providing both graphical and numerical methods to assess the performance of the algorithm. Particularly useful is the construction of a trace plot – a time series graph of a sampled parameter or statistic for each sample from the MCMC plotted against the iteration number.

Burn-in. Since the analysis starts at an arbitrary initial value, the posterior density is likely to be low and it may take a while for the chain to sample from the highest density regions of the posterior. This initial period is called the burn-in, and it is common to discard these samples from the analysis as it makes samples closer to the stationary distribution and less dependent on the starting point. Therefore the duration of burn-in used varies, depending on how close the initial value is to the stationary distribution. As the chain reaches the stationary distribution, the posterior density and parameter estimates begin to plateau, shown by the trace appearing horizontal. Comparing the results from multiple MCMC runs (particularly from different initial values) helps reveal whether chains are converging on the same posterior distribution.

Thinning. It is usual to only save a fraction of the samples from a long MCMC chain at regular intervals, in order to save on disk space and computer memory. Although this process, known as thinning, wastes some information, it is relatively efficient because there is a natural dependency between adjacent samples from the chain. If there is high dependency across a chain of draws, through correlation or periodicity, then multiple samples add very little to a single sample. An effective sample size (ESS) can be calculated by taking the chain length minus the period of burn in, and dividing it by the autocorrelation time given by the average number of states needed to separate two samples for them to be considered independent. Low ESS values suggest that there is high correlation between samples, and the chain should be run for longer.

2.4 Model Selection

Model selection is a key part of statistical inference and scientific hypothesis testing. Often there are multiple competing hypotheses, and some assessment is required to choose which of these models is 'best'. In general, models with more parameters will more closely fit the data. However, this can be computationally intensive, and with more parameters estimation tends to decrease in accuracy. As such, which model is considered best comes from a balance of the scientific relevance (the biological plausibility), the goodness of fit, and complexity (Steel 2005).

2.4.1 Bayes Factor Tests

The Bayes factor (BF) gives the relative evidence in favour of one model compared to another in the Bayesian setting, similar to the role of the likelihood ratio test in

frequentist statistics. A quantity known as the marginal likelihood, $p(X|M)$, measures the average fit of a model M to the data X, averaged over all possible values of the parameters. The Bayes' factor is defined as the ratio of the marginal likelihoods of the two models (Kass and Raftery 1995):

$$\mathrm{BF} = \frac{p(X|M_2)}{p(X|M_1)}. \tag{2.7}$$

A Bayes factor significantly greater than one implies that there is more support for M_2 than M_1. Whilst there is no formalism equivalent to a classical hypothesis test for interpreting the Bayes factor, Jeffreys (1961) and Kass and Raftery (1995) provide a guide to its interpretation, suggesting that Bayes Factors greater than 20 show strong evidence in favour of M_2.

2.5 Coalescent Theory

The coalescent model describes the genealogy of a sample of genes under a simplifying set of assumptions known at the standard neutral model (Kingman 1982a; Kingman 1982b). It has theoretical and practical advantages over other approaches that describe the genealogy of an entire population, such as the Wright-Fisher (Fisher 1930; Wright 1931) and Moran (Moran 1957) models. Whereas these traditional approaches describe the evolution of a population forwards in time, the coalescent is defined backwards in time, as a retrospective description of the genealogy of a finite sample. The main advantage of the coalescent over the Wright-Fisher and Moran models is that only the history of the sample need be considered. This greatly simplifies the model and makes it much more tractable for statistical inference. It is also very flexible and readily adaptable

for various deviations from the standard assumptions (Nordborg 2008), of which those relevant to this thesis are explained in the following sections.

2.5.1 The Standard Coalescent

First, consider a finite population containing a total of *N* individuals. The standard neutral model upon which the coalescent is based makes three core assumptions:

- the population is panmictic,
- it has a constant population size through time, and
- it is not subject to selection.

The retrospective view that the coalescent takes begins by considering the relationship between individuals in the current population and their parents in the previous generation. Lineages (which describe the ancestry or relatedness of individuals through from parent to offspring) can be traced from the current generation to the previous one. When an individual is the sole offspring of a parent, the lineage extends backwards one generation. When there are two or more individuals with the same parent, their lineages merge or coalesce in the previous generation.

For a locus with ploidy P, there are PN copies of the gene in the population. Accordingly, the probability that copies of the gene in the current generation have the same parent is, under the standard neutral model, $1/PN$. In other words, the probability that two lineages coalesce in the previous generation is $1/PN$ and the probability that they do not coalesce in any of the preceding PNt generations is $(1 - 1/PN)^{PNt}$. As the population size gets large (formally as $N \to \infty$) this probability is well approximated by

the asymptotic limit e^{-t}. This indicates that the waiting time (in units of PN generations) until a pair of lineages coalesce can be modelled by an exponential distribution with rate equal to one. Assuming further that coalescent events are rare and can therefore be deemed independent, the total coalescence rate when there are k lineages is given by $k(k-1)/2$. Note that it is standard practice, when using the coalescent, to scale time in units of PN generations. Since the organisms covered in this thesis are haploid, $P = 1$ will be taken for what follows.

Now, consider a sample of k genes sampled in the present. The coalescent can be described in terms of the following generative process. Initially there are $k(k-1)/2$ possible coalescent events, with each pair of lineages equally likely to be involved in the coalescence. The waiting time to this coalescence in exponentially distributed with rate equal to $k(k-1)/2$. The process then continues with one fewer lineage until only a single lineage, the most recent common ancestor of the sample, remains. The coalescent is thus computationally efficient for simulation purposes, as for a sample of size n only $n-1$ independent exponential random variables need to be sampled alongside randomly selecting a pair of lineages to coalesce at each event. On average, as the number of lineages decreases, the waiting time between coalescence events increases. This leads to genealogies with coalescence events occurring predominantly near the tips, and longer waiting times towards the root of the tree.

2.5.2 Effective Population Size

It is inevitable that a simplified mathematical model such as the coalescent cannot fully accurately describe a natural population; it is necessary to balance the biological realism of any model with practical concerns such as the tractability of the model for mathematical analysis and computational inference. However, two assumptions of the standard coalescent model that are frequently likely to be unrealistic are random mating and a constant sized population. Remarkably, many extensions to the standard neutral model exert their effect through a straightforward shift in the population size. For this reason, the notion of an *effective* population size – denoted N_e – is broadly applicable and widely used.

The effective population size is a parameter that allows the coalescent model to be applied to real world problems – essentially mapping the features of a more complex natural population to a simpler and well formulated model with known results and well understood behaviour (Sjödin et al. 2005). If N is considered the census population size, then N_e is the size of an idealized population that would produce similar dynamics in the simplified model. One way in which population dynamics are quantified is through genetic diversity. In the standard coalescent, genetic diversity is proportional to the population size. This relationship allows N_e to be calculated by finding the ideal population size under the standard neutral model that produces the same diversity as observed in the natural population. Two examples of using the effective population size are given below.

Variation in Reproductive Success. As a consequence of the assumptions underlying the standard neutral model, the number of offspring of each potential parent in one generation is Poisson distributed with mean one (so that the population size stays constant), and variance also equal to one. Now, in real life there will often be greater variance in reproductive success than expected under the standard assumption that every potential parent is equally fit. This is likely even if that variation in fitness is not inherited, for example because of differences in access to resources. Then, the variance might be given by some value of σ^2, where $1 < \sigma^2$. So long as fitness differences are not heritable, this situation converges to the standard coalescent with time scaled by the variance, so that the effective population size is given by $N_e = N/\sigma^2$. Intuitively, this happens because higher variance in reproductive success results in a smaller gene pool contributing to the next generation. The effective population size is therefore smaller than the census size, coalescence occurs relatively faster, and genetic drift is seen to operate more strongly (Nordborg 2008).

Strong Migration Limit. Another useful result concerning the effective population size arises from structured populations with strong migration between sub-populations. It is worth mentioning briefly here, as similar principles will be revisited in Section 3.3.2 when looking at a coalescent model for metapopulations. Suppose the population has some geographical structure, so that it is divided into smaller sub-populations with migration occurring between them. The strong migration limit arises when the sub-populations are relatively large so that coalescent events are rare compared to migration of lineages between sub-populations. With sufficiently strong migration, lineages

migrate at such a high frequency the population appears to be essentially panmictic (Wakeley 2004a). Thus the waiting time until two lineages coalesce depends primarily on the probability that they happen to be in the same place at the same time. Further, the location of a lineage at any time is effectively independent of its location at any other time. The population structure therefore becomes unimportant except that it decelerates the rate of coalescence by some factor α that depends on the number of sub-populations and the migration rates between them. The structured coalescent with strong migration therefore converges on the standard coalescent model with effective population size $N_e = N/\alpha$ (Nordborg 2008).

2.5.3 Demographic Change

As noted above, in the coalescent model, the rate at which lineages coalesce is determined by the population size. If the population is large, then the probability of two lineages having the same parent is small, and thus the rate of coalescence is slower than in a small population. Fluctuations in the population size can be modelled as a standard coalescent process with a time-varying coalescence rate that depends on the population size (Griffiths and Tavaré 1994; Donnelly and Tavaré 1995).

At any time t, the rate of coalescent relation to that at time 0 can be defined as

$$\lambda(t) = \frac{N_0}{N_t}, \tag{2.8}$$

where N_0 and N_t are the number of individuals in the population at time 0 and t respectively. Then the pairwise cumulative rate of coalescence, $\Lambda(t)$, from the present to time t is defined as

$$\Lambda(t) = \int_0^t \lambda(u)du. \tag{2.9}$$

The time until a pair of lineages coalesces becomes an inhomogeneous waiting time process with probability density function

$$\mathrm{p}(t) = \lambda(t)\exp\big(-\Lambda(t)\big). \tag{2.10}$$

More generally, for a sample of k lineages, probability density function for the time to the next coalescence becomes

$$\mathrm{p}(t) = \binom{k}{2}\lambda(t)\exp\left(-\binom{k}{2}\Lambda(t)\right). \tag{2.11}$$

Due to the changing population size, the waiting times until coalescence are inhomogeneous, and are related to the standard coalescent using a non-linear rate of coalescence scaled according to $\Lambda(t)$. This means that simulation of a coalescent model with varying population size is straightforward, only requiring a transformation of the times simulated under a standard coalescent model (Hein et al. 2005). This transformation is given by $t' = \Lambda^{-1}(t)$, where t' is the waiting time required for the changing population size, and t is simulated under a demographically stable population of size N_0. It is often possible to find $\Lambda^{-1}(t)$ in closed form; however if the transformation is more complicated then numerical methods may be used. This method is used in Section 3.3.6 to implement the coalescent SIR model in BEAST.

2.5.4 Mutation

Mutation is simple to incorporate into the coalescent because of one of the key assumptions: selective neutrality. Under selective neutrality, the population birth-death

process is unaffected by the mutational state, and therefore the mutation process is independent of the genealogical process. This is important for simulating data under the coalescent because it means that the mutation process can be superimposed on a previously simulated coalescent tree (Hein et al. 2005).

Assuming a constant mutation rate, the mutations that occur on the tree follow a Poisson process with rate $\theta/2$, where θ is the scaled mutation rate $2N\mu$ and μ is the average number of mutations differing from offspring to parent. This means that the expected number of mutations will increase linearly with branch length, and between a pair of individuals equals θ. The frequency of coalescence events tends to decrease towards the root of the tree, when only a few lineages remain, such that mutations occurring within these lineages tend to dominate the sequence divergence much more than mutations occurring near to the tips of the tree.

2.6 Modelling Evolution

A model in which mutations are assumed to occur independently and at a constant rate, as in the previous section, and each mutation alters a previously unmutated site, is known as an infinite sites model (Kimura 1969). The infinite sites model may be appropriate when the locus under investigation is composed of many nucleotides, and mutation is sufficiently rare so at any single nucleotide in the sequence, at most only one mutation is likely to occur. These assumptions make the model analytically tractable, but limit it in two distinct ways. First, the model assumes that repeat mutations at a single site in the genome are not allowed. Whilst this may be reasonable over a short

timescale with a slow mutation rate, the model is limited and its application to datasets with multiple alleles per site is problematic, a situation regularly encountered for viral pathogens with relatively small genomes and fallible replication processes. Secondly, it does not take into account the current allelic state, and thus does not account for how this might affect the rate of mutation. This section looks at more elaborate nucleotide substitution models, with examples of some of the most widely used models in phylogenetics.

2.6.1 Nucleotide Substitution Models

Nucleotides are split into two biochemical classes – the purines (adenosine and guanine) and the pyramidines (cytosine, thymine and uracil). Mutation events can be classed into two types (Figure 2.6.1); transitions are mutation events that do not change the nucleotide class type (so, from adenosine to guanine, for example), whereas transitions are mutations events that do change the nucleotide class (such as adenine to guanine).

In the simplest model of nucleotide evolution, the Jukes-Cantor model (Jukes and Cantor 1969), the four nucleotides are assumed equally common and mutation rates between any pair are the same. Multiple mutations can, however, occur at the same position. Empirically, however, transitions are observed to occur more often than transversions, despite there being twice as many possible transversions. The Kimura two-parameter model (Kimura 1980) is the simplest nucleotide model that takes this phenomenon into account via the transition:transversion ratio, usually represented by the parameter κ.

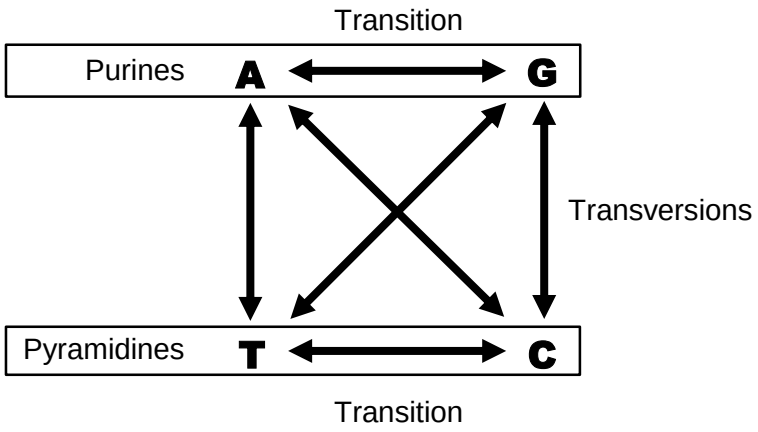


Figure 2.6.1: Classification of nucleotide substitutions. There are six possible substitutions. Transitions occur when a nucleotide is replaced with one of the same class, and transversions occur when a nucleotide is substituted with one of a different type. For RNA substitutions, thymine (T) is replaced with uracil (U).

The process of nucleotide substitution (when one nucleotide mutates to, or is substituted, by another) can be represented mathematically using a Markov process with instantaneous rate matrix, **Q** (Strimmer and von Haeseler 2009). The Markov property arises naturally from an assumption that the rate of substitution from one nucleotide to another is only dependant on its current state, and independent of any changes that have occurred at the site previously. By convention, the rate matrix is ordered alphabetically according to nucleotides for row and columns (A, C, G, T), so that element Q_{ij} refers to the rate of a nucleotide i being replaced by j at a given site.

In a general time-reversible model, the non-diagonal elements are given by the product of the mean substitution rate (μ), the relative rate of nucleotide i mutating to j (by construction, $R_{ij} = R_{ji}$) and the frequency of the nucleotide to which the current one is mutating (π_j). The diagonal of **Q** represents the rate at which the current state mutates to an (observably) different nucleotide, and is easily calculated since the sum of each

row must equal zero. It is also assumed that the relative frequencies of the four nucleotides are at equilibrium, and that $\sum \pi_i = 1$. The probability of a given nucleotide mutating to any other during a time t is then given by the transition probability matrix:

$$\mathbf{P}(t) = \exp(\mathbf{Q}t). \tag{2.12}$$

While the most general model of substitution of this form is the general time reversible (GTR) model (Tavaré 1986), it may be simplified by placing restrictions on the differing rates of transitions and transversions. In the TN93 model, Tamura and Nei (1993) define the parameters κ and γ as the ratio of transitions to transversions, and ratio of purine to pyramidine transition rates, respectively. The HKY85 model (Hasegawa et al. 1985) is a simplification of this, with $\gamma = 1$, and the Kimura 2-parameter model (Kimura 1980), K80, further reduces the number of parameters by assuming that all of the nucleotide frequencies are equal. As mentioned previously, the simplest model of substitution is the Jukes-Cantor model (Jukes and Cantor 1969), where there are equal proportions of each nucleotide base, and all possible substitutions are deemed to be equally likely.

When parameter estimation is required, there is the usual trade-off between the biological complexity of a particular model, and computational time required. The matrix $\mathbf{Q}$ must be exponentiated to obtain the transition probability matrix $\mathbf{P}(t)$, and while this is possible analytically for the HKY85 model, it is not possible for the GTR model. In addition, more complex models can lead to large variation in the estimated number of substitutions per site (Rzhetsky and Nei 1995), though this is less of an issue with whole genome datasets. Nei and Kumar (2000) show that up to 0.5 expected

substitutions per site, the Tamura (an extension of the K80 model allowing for high or low GC contents by Tamura (1992)), K80 and Jukes-Cantor models all give a reasonable approximation when the nucleotide evolution is known to follow the TN93 model, and thus support the use of the simplest model for samples of closely related sequences.

More generally, however, software exists to assess which model offers the best fit according to the number of parameters in the model. ModelTest (Posada and Crandall 1998) takes input in the form of the log-likelihood for each model such as those output from the tree-building software PAUP* (Swofford 2003), and then uses hierarchical likelihood ratio tests alongside the Akaike Information Criterion (Akaike 1974) and Bayesian Information Criterion (Schwarz 1978) to rank the models according to best fit.

2.6.2 Rate Heterogeneity

So far, it has been assumed that the rate of substitution is constant over all sites. This assumption can be relaxed to allow sites to mutate at different rates – for example in a codon model where the second and third positions mutate at a faster rate than the first. This heterogeneity is often represented by the gamma distribution with mean equal to μ and a flexible parameter α for the shape. When α is less than one, the gamma distribution is L-shaped. This is equivalent to strong rate variation, where some sites have high rates of substitution, but the majority have a low rate. For $\alpha > 1$, the distribution of rates around the mean becomes tighter, and as α tends to infinity, the model converges to the constant rate substitution model. Using a continuous gamma distribution of rates is computationally intensive, so it is standard to assume a discrete

gamma distribution with four to eight rate categories of rates as a reasonable approximation (Yang 1994).

2.7 Recombination

A recombinant genome is one in which different parts of the genome have different ancestral histories, i.e. different genealogies. This means that the genome is made up of a mosaic of sources with varying evolutionary histories. Recombination in a sample means that the evolutionary tree is not identical for all sections of a nucleotide sequence, and thus a single tree cannot represent the ancestral relationships between samples. This conflict in ancestry in an alignment is known as a genetic incompatibility (Figure 2.7.1), and can be identified with Hudson's four gamete test (Hudson and Kaplan 1985).

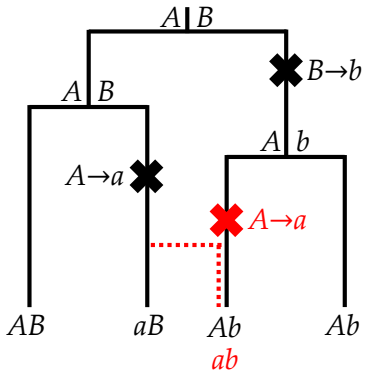


Figure 2.7.1: Diagram illustrating genetic incompatibility. For two loci A and B, a maximum of two mutations (depicted with crosses) are expected under the infinite sites model, and thus a maximum of three haplotypes (AB, Ab and aB) can be obtained (black). The fourth haplotype ab is only possible through recombination (dashed red line), or recurrent mutation (red cross).

If there are two loci, one with alleles A and a, and the other with alleles B and b, there are four possible combinations (or haplotypes): AB, Ab, aB, ab. Under the infinite sites model (Kimura 1969), it is impossible to observe all four haplotypes in a sample of sequences. Therefore an incompatibility is either the result of recombination having occurred or more than one mutation taking place at a single site.

2.8 Reconstructing Phylogenies

In this thesis, PhyML (Guindon et al. 2010) was used to estimate phylogenies by maximum likelihood, and BEAST (Drummond et al. 2012) was used for Bayesian phylogenetic inference.

2.8.1 Maximum Likelihood Phylogenetic Methods

Maximum likelihood methods involve finding the tree and evolutionary parameters that maximise the probability of observing the sequence data. The likelihood of a tree topology and branch lengths given the observed sequence data can be found using Felsenstein's pruning algorithm (Felsenstein 1981). This starts at the tips of the tree, and works along the tree towards the root calculating the probability of the sequences given the tree so far, and multiplying across sites in the sequences. The branch lengths can be optimised using numerical computation techniques. The main problem is quickly exploring the tree space, since for large datasets there are a vast number of possible topologies.

Various moves for exploring phylogenetic tree space have been developed. Full-tree rearrangement starts with an initial tree, and then generates a 'neighbourhood' of trees around this (Schmidt and von Haeseler 2009). For each tree in the neighbourhood, the one with the highest likelihood is found, and taken as the starting tree in the next iteration. If no trees in the neighbourhood are better than the current one, then the current tree is taken to be the optimal one. There are three main rearrangement methods for obtaining the neighbourhood of trees: nearest neighbour interchange, sub-tree pruning and regrafting, and tree-bisections and reconnection (Felsenstein 2004). PhyML 3.0 uses sub-tree pruning and regrafting alongside parsimony methods to find the tree which maximises the likelihood function (Guindon et al. 2010).

2.8.2 Bayesian Phylogenetic Methods

Phylogenetic inference in the Bayesian framework involves finding the posterior distribution for a set of evolutionary parameters given the sequence data, where parameters include the phylogenetic tree topology and branch lengths, and elements of the substitution model. The parameters are sampled using the Metropolis-Hastings MCMC algorithm (Section 2.3.1). To accelerate convergence, a starting tree is usually proposed using a fast tree building method such as UPGMA (Sokal and Michener 1958). Once the MCMC run has reached convergence, the post burn-in samples are taken as an approximation to the posterior distribution (Section 2.3.32.3.3). Bayesian methods have the advantage that by sampling from a posterior distribution, there is no need to condition on a single tree when making inference, as in the distance based and maximum likelihood methods. This means that all of the uncertainty in the topology,

branches, and evolutionary models can be taken into account. Methods for summarising a Bayesian analysis are given in Section 2.8.6.

One of the most widely used implementations of Bayesian phylogenetic inference is BEAST (Bayesian Evolutionary Analysis by Sampling Trees). BEAST takes input in the form of an XML file that describes the data to be analysed, the model to use, information about the MCMC process (such as the number of iterations, and proposal distribution), and the desired output (Drummond et al. 2012). For commonly used models, the easiest way to format this file is to use the graphical user interface, BEAUTi, which is distributed in BEAST. More complicated combinations of models must be edited in to the XML file directly. Other Bayesian phylogenetic approaches such as MrBayes (Huelsenbeck and Ronquist 2001), LAMARC (Kuhner 2006) and BATWING (Wilson et al. 2003) are also in wide usage. However, in this thesis I used BEAST because it supports a wide range of models, and is open to development so that new methodologies are readily incorporated.

2.8.3 Relaxed Clocks

An alternative to the strict molecular clock described in is to assume the evolutionary rates are independent for each individual branch. Unrooted trees can be interpreted in these terms since branch lengths are estimated in units of the expected evolutionary change per site, a compound quantity representing the product of the branch length in years and the yearly substitution rate (Felsenstein 1981). Whilst methods are widely used to infer unrooted phylogenies (Huelsenbeck and Ronquist 2001; Swofford 2003;

Felsenstein 2005), a limitation is that it is not possible to individually estimate the molecular rate and branch length if they are both allowed to vary independently over branches. Moreover, the position of the root cannot be identified without assuming a suitable out-group. One source of information allowing the deconvolution of these quantities is the sampling time, which can be used to estimate rooted trees using software such as BEAST.

BEAST implements a number of molecular clock models sitting between the two extremes represented by the strict clock and unrooted model. Local molecular clocks estimate different rates for user-defined clades within the tree (Hasegawa et al. 1989; Rambaut and Bromham 1998; Yoder and Yang 2000). However, such methods are clearly not ideal for large numbers of isolates, or topologies with large amounts of uncertainty – and thus their use is restricted to situations where certain taxa are known to have rates different to the rest of the tree. Drummond et al. (2006) introduced a new class of relaxed clock models, where the evolutionary rate is allowed to vary among branches according to some distribution defined by the user. Previous attempts at relaxing the clock assumed that closely related taxa have similar evolutionary rates. This is known as the autocorrelated relaxed clock, and assumes that the rate of evolution on a branch is dependent on the rate of the parental branch immediately before it (Thorne et al. 1998; Aris-Brosou and Yang 2002). Drummond et al. proposed an uncorrelated relaxed clock that does not make this assumption, so that rates on closely related branches do not have to be correlated. The uncorrelated relaxed clock is currently

implemented in BEAST under the log-normal and exponential distributions (Drummond and Rambaut 2007).

Further, Drummond et al. (2006) show that assuming a strict clock for data in which there is heterogeneity in rates across branches gives generally poor results, with the true rate included in the 95% HPD only 3-11% of the time. However, uncorrelated relaxed clock models still performed well when the true clock model had a fixed rate across the tree. In BEAST, it is possible to use the Bayes factor to identify the best fitting of two or more different clock models (Li and Drummond 2012).

2.8.4 Serially Dated Samples

A measurably evolving population is one that has accumulated a significant number of observable mutations between dated samples (Drummond et al. 2003b). This definition covers both species which have a slow evolutionary rate, but for which ancient samples have been obtained and thus cover a long period of time (Barnes et al. 2002; Shapiro et al. 2004), and those that evolve rapidly such as RNA viruses which adapt quickly to their environment and thus are likely to have measurable diversity between sampled isolates, even those collected close in time.

In the clinical setting, it is highly likely that samples have been collected over a period of time as isolates are collated during the course of an outbreak, or routinely as people enter the hospital, and these temporal data can be exploited to give insights into the demographic history and dynamics of the population (Drummond et al. 2003a). The use

of dated tips in BEAST allows the rate of substitution (μ) to be separated from the evolutionary time (θ), thus allowing the time scale of the phylogeny to be estimated in terms of real time (e.g. the substitutions per site per year) as opposed to just the expected number of substitutions per site (Drummond et al. 2002). The timescale can then be used to date specific events in the ancestry of a data set, for example the time to the most recent common ancestor, and to compare these to known historical occurrences, such as the use of certain treatments.

2.8.5 Evaluating Uncertainty in a Tree

Bootstrapping (Efron 1979; Efron and Tibshirani 1994) is commonly used in phylogenetics to evaluate the uncertainty in a point-estimate of a phylogeny (Felsenstein 1985), especially when using maximum likelihood and distance based methods. In the case of phylogenetic trees, bootstrapping usually takes the form of resampling from the columns of the original sequence alignment with replacement to obtain a new dataset the same size as the original. Given an alignment where rows represent isolates, and columns are sites in the genome, the columns are randomly resampled with replacement to obtain a new alignment of sites with the same length, making the assumption that sites in the alignment have evolved independently. The new alignment is known as a bootstrap sample, and typically a bootstrap analysis will involve up to 1000 such resamples of the original data (Pattengale et al. (2010) discuss how many bootstraps should be performed for accuracy versus computational cost). Each bootstrap sample is used to construct a bootstrap tree using the same method as for the original tree, which can then be compared to the phylogeny constructed from the original data –

for example to investigate the proportion of trees in which the same clade is identified. Metrics for the comparison of tree topologies included the Robinson-Foulds metric, or symmetric difference, (Robinson and Foulds 1981), and Branch Score Distance (Kuhner and Felsenstein 1994), both of which are implemented in TreeDist in PHYLIP (Felsenstein 2005). Theoretically, the bootstrap samples capture the variation that would have been obtained if it had been possible to sample many such evolving data sets, and therefore is an estimate of the error in the original phylogeny.

2.8.6 Summarising a Posterior Distribution of Trees

Unlike in maximum likelihood methods where a point estimate of the tree is obtained, Bayesian analyses produce a posterior distribution of phylogenies. In most cases, it is desirable to summarise this distribution for visualisation and to make inferences about, for example, divergence times between isolates. There are several ways to summarise a Bayesian phylogenetic analysis and obtain the posterior probabilities of ancestral relationships. Majority consensus methods construct the summary tree so that it contains all the clades that occur in 50% or more of the posterior samples (Margush and McNorris 1981). The remaining clades that are not well supported are given as collapsed nodes (polytomies) that have more than two branches descending from them. The extended majority consensus tree, also known as the greedy consensus tree, takes this further, and gives a fully resolved tree whereby the remaining clades are added in order of decreasing probability, as long as they are in agreement with the previously resolved clades (Bryant 2003). The limitation of these methods is that it is possible for the resulting consensus tree to have a topology that was never sampled in the posterior, or

only sampled at low frequency (Cranston and Rannala 2007; Drummond 2010). Thus whilst some clades will have high support, the topology may not best represent a biologically viable or realistic scenario.

Alternatively, a maximum *a posteriori* tree can be found. Drummond (2010) notes that the definition of the maximum *a posteriori* tree has previously been defined as the tree in the posterior samples associated with the highest density, but that this is problematic as the high posterior probability density may be due to the fit of the branch lengths rather than the topology. Instead, he suggests that the maximum *a posteriori* tree should be the topology that has the highest posterior density, averaged over the other parameters. For well resolved sets of data with few taxa, this is the topology which is most sampled in the MCMC analysis. However, with large datasets it is likely that sampled trees have unique topologies and thus methods exist to estimate the probability of a particular tree topology. The maximum clade credibility (MCC) tree is defined as the one that maximises the product of the clade posterior probabilities, and the maximum credibility tree is the one that maximises the sum of the clade posterior probabilities.

In this thesis, I have used TreeAnnotator (distributed with BEAST) to summarise a sample of trees obtained using BEAST, and obtain posterior statistics such as credible intervals around the node heights. The summary phylogenies were then visualised and annotated using Figtree (available from: http://tree.bio.ed.ac.uk/software/figtree/).

Chapter 3: Coalescent Inference for Infectious Disease

3.1 Introduction

Hepatitis C virus (HCV) is one of the most intensively studied organisms in the context of joint evolutionary and epidemiological inference. Indeed, HCV was the first pathogen for which coalescent analyses were formally used to investigate epidemic behaviour (Pybus et al. 2001), and it has since stayed in the spotlight both due to its medical importance (see Section 3.2) and amenability to genetic analysis. Previously, coalescent inference has been used to date the emergence of HCV (for example, by Verbeeck et al. (2006), Magiorkinis et al. (2009), Pouillot et al. (2008) and Pybus et al. (2009)) , and to provide evidence for the roles of poor medical practice (Pybus et al. 2003) and drug use in transmission (Pybus et al. 2005).

However, the underpinnings of the most widely used coalescent models for epidemiological inference have recently been called into question, casting doubt on their interpretation. Traditionally it has been assumed that the effective population size is directly proportional to the prevalence of infection, allowing changes in the population size to be used in the estimation of R_0 (Pybus et al. 2001; Pybus et al. 2003). Theoretical work has shown that this is only valid at dynamic equilibrium (Koelle and Rasmussen 2012) and not more generally (Volz et al. 2009; Frost and Volz 2010).

This chapter begins with an overview of the epidemiology and evolution of hepatitis C, for context. I then present robust population genetic inference for compartmental

models in epidemiology, using a metapopulation model to represent the pathogen population, as described in Dearlove and Wilson (2013). This coalescent epidemiological approach will then be utilised to investigate the hypothesis that underlying epidemiological processes (such as the growth rate) and demographics of the host population are driving the genetic diversity of HCV epidemics. This will be tested through conducting a meta-analysis of previously published HCV datasets, enabling a direct comparison of the different dynamics of epidemics from across the world, including the effect of subtype. Such a modelling approach offers a better view to the long-term epidemiology of disease across multiple epidemics, adding an extra facet to the global understanding of a highly complex disease.

3.2 Hepatitis C Virus

A member of the genus *Hepacivirus* in the *Flaviviridae* family, HCV is one of the major causes of progressive liver disease, leading to the virus being among the most frequent reasons for liver transplantation. It affects over 150 million people worldwide, with an estimated three to four million new cases and 350,000 deaths due to HCV-related liver disease each year (Kim et al. 2013). High prevalence of disease is found in the countries of Africa, Asia and South America (Shepard et al. 2005); the highest reported prevalence of HCV antibody is in Egypt at 14.7% (El-Zanaty and Way 2009). However, HCV is not restricted to these areas and the developing world; it is the most common chronic infection in the US affecting up to 3 million individuals (Kim 2002), causing an estimated 8,000 to 10,000 deaths annually (Centers for Disease Control and Prevention 1998) and representing an annual cost of $744 million (Kim et al. 2002). The impact of

disease is only set to rise in the next twenty to thirty years, and thus will continue to be a huge global burden on public health.

HCV was first cloned and identified in 1989, after the discovery of a hepatitis that was of neither A nor B type (Choo et al. 1989). So far, the virus has not been found to naturally infect species other than humans, although chimpanzees can be experimentally infected with the virus and do show symptoms (Bassett et al. 1998). Attempts to study the disease were limited for a number of years by the inability to culture the virus *in vitro*, but this hurdle was overcome in 2005, when a complete infectious cell-culture system was established (Lindenbach et al. 2005; Wakita et al. 2005). Work continues to establish a mouse model to study HCV entry and replication in greater detail (Mercer et al. 2001; Bissig et al. 2010; Dorner et al. 2011).

3.2.1 Structure and Evolution

HCV is an enveloped virus, thought to have an outer diameter of around 55nm (Wakita et al. 2005). It has a positive-strand ribonucleic acid (RNA) genome of around 9.6kb, representing a single open reading frame (ORF) flanked by untranslated regions (UTRs) (Fusco and Chung 2012). The ORF is approximately 3,300 codons long, and encodes a polyprotein made up of three structural proteins (core, E1, and E2) and seven non-structural proteins (Figure 3.2.1). It has been suggested that the envelope glycoproteins E1 and E2 are attached to a double-layer lipid membrane surrounding a nucleocapsid containing copies of the core protein and the RNA genome (Moradpour et al. 2007). E2 is made up of three hypervariable regions shown to be under strong selective pressure,

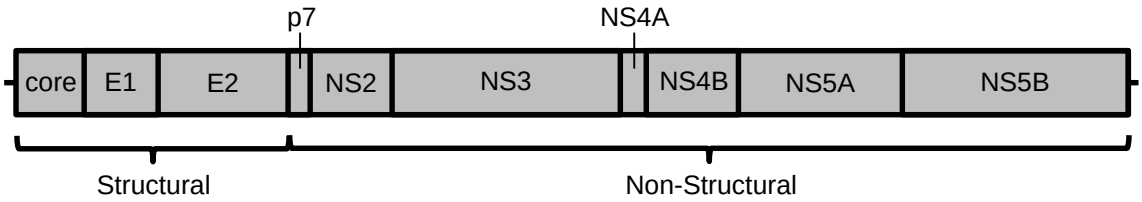


Figure 3.2.1: Structure of hepatitis C genome. The hepatitis C genome is made up of a single open reading frame, split into ten proteins. The structural proteins include the core, and two envelope proteins (E1 and E2). The RNA-dependant RNA polymerase is found within NS5B. Adapted from Figure 3 in Moradpour et al. (2007).

which is unsurprising given the role the glycoprotein has in the immune response (Troesch et al. 2006). Of the seven non-structural proteins, NS3/4A encodes (amongst other enzymes) a serine protease responsible for downstream cleaving (Yao et al. 1999; Wölk et al. 2000), and NS5B encodes the RNA-dependant RNA polymerase (RdRp) required for replication (Behrens et al. 1996). Due to their role in virus replication and assembly, these proteins have been among the main focal points in drug development (Lamarre et al. 2003; Powers et al. 2006).

HCV is split into seven subgroups, labelled 1 through 7, and is further classified by subtypes denoted with lower case letters, for example, 1a (Nakano et al. 2012). The virus is known to be one of the most diverse human viral pathogens, exhibiting more than 30% nucleotide divergence over the genome between the six defined subgroups. Further, subtypes within a subgroup show 20-25% nucleotide diversity (Simmonds et al. 2005). Despite the global coverage of HCV, there remains an extensive difference in the global distribution of subtypes. Genotypes 1a, 1b and 3a are circulated widely in the developed world, and constitute most of the infections seen in the clinical setting (Simmonds 2004). The other subtypes are more distinct in their geographical coverage.

Infections in West Africa are most often found to be of subgroup 2 (Jeannel et al. 1998; Candotti et al. 2003), whereas subgroup 4 is associated with Central Africa and the Middle East (Fretz et al. 1995; Chamberlain et al. 1997; Ndjomou et al. 2003), and subgroups 3 and 6 with East Asia (Pybus et al. 2009).

The high diversity in subtypes is in part due to the mechanism of replication. The RNA-dependent RNA polymerase allows rapid replication but due to a lack of proof-reading, is also error prone (Ortín and Parra 2006). The estimated mutation rate varies widely according to study, region of the genome, and method used; a brief literature review yields rates from 3×10^{-4} to 1.0×10^{-3} substitutions per site per year (Tanaka et al. 2002; Pybus et al. 2005; Magiorkinis et al. 2009; Gray et al. 2011). Together with a generally high rate of viral production, between 10^{10} to 10^{12} virions per day, and an estimated half-life in serum of two to three hours, this can result in a rapid accumulation of mutations (Neumann et al. 1998). Indeed, it has been suggested that HCV circulates as a population of closely related virions, that some refer to as "quasi"-species (Martell et al. 1992), although this term does not imply levels of diversity on a par with that of the species, or even with that of subtypes (Holmes 2010).

3.2.2 Transmission

There are a number of risk factors attributed to the spread of HCV, including the increased use of blood products (such as transfusion), tissue transplantation (particularly the liver), haemodialysis, and non-sterile needle use (Pereira et al. 1991; Hauri et al. 2004; Pybus et al. 2005; Goedert et al. 2007; Nelson et al. 2011). Before routine

screening was introduced in developed countries in the 1990s, it was estimated that one in every 50 units of blood transmitted Hepatitis C (Alter et al. 1981; Colombo et al. 1987). Now, the figure is estimated to be less than one in 250,000 (Allain 2003), with the predominant cause of infection shifting to injecting drug use. This is evidenced by a distinction in circulating subtypes. In the US and Europe, subtype 1a is generally found in younger individuals, with injecting drug use as the main risk factor, whereas subtype 1b is more commonly found in older individuals who have had a blood transfusion (Pawlotsky et al. 1995).

However, in the developing world healthcare associated transmission remains high. Between 2001 and 2002, six million units of blood were not screened for blood-borne infectious diseases, and 68 out of 131 countries were not screening all units of blood (World Health Organisation 2004; Prati 2006). By 2004, this had reduced to 37 countries out of 147 that responded, but this still remains a significant risk (World Health Organisation 2008). Contaminated injections also pose a fundamental risk, with reuse of non-sterile syringes common, and many injections being unnecessary due to the possibility of medication easily being taken in other forms (Simonsen et al. 1999; Hutin et al. 2003).

Other modes of transmission include tattoos, especially those done in non-professional settings (Jafari et al. 2010), community barbershops (Bari et al. 2001), and vertical transmission from mother to baby during pregnancy (Lam et al. 2010). Unlike hepatitis B virus and human immunodeficiency virus, transmission of HCV through sexual

contact is much less efficient than other routes and therefore relatively rare, with transmission between long term sexual partners more likely to be due to common exposure to other risk factors (Stroffolini et al. 2001; Vandelli et al. 2004).

3.2.3 Clinical Infection and Diagnosis

HCV infection tends to manifest in two ways: either appearing self-limited with spontaneous viral clearance, or persisting to chronic disease. After infection, the acute phase of disease has an incubation period of up to 20 weeks. After this time, clinical symptoms, if they occur, tend to be unspecific including fatigue, abdominal pain, nausea and jaundice (Hoofnagle 1997). The generality of these symptoms means that few acute infections are reported, limiting the study of spontaneous viral clearance. However, epidemiological cohort studies have shown that infection is chronic (having a duration of at least 6 months) in 75-85% of cases (Chen and Morgan 2006).

Chronic HCV infection is often asymptomatic for the first several years of infection, though can be associated with fatigue (Poynard et al. 2002). Later in the course of infection, there may be inflammation of the liver, which can proceed to liver fibrosis (an excess of fibrous tissue due to repair) and cirrhosis (scarring). Amongst patients with HCV infection and cirrhosis, there is a 1-8% chance of developing hepatocellular carcinoma (HCC), a type of liver cancer. The risk of HCC is increased in those who are male, were infected at an older age, have coinfection with hepatitis B or HIV, or consume high amounts of alcohol (Fassio 2010). Studies on the progression of chronic infection are limited, as it requires decades of follow up (Alter and Seeff 2000).

Since it is common for patients to remain asymptomatic for years, many only discover that an infection has taken place many years later during health screening, for example for a blood transfusion (Contreras et al. 2010). Initial diagnosis is usually by either serological assay such as the immunoblot, which detects four antibodies to HCV using nitro-cellulose strips coated in viral antigen, or molecular assay, which detects any viral nucleic acid that is present (van der Poel et al. 1991; Ghany et al. 2009). Quantitative molecular assays, including polymerase chain reaction (PCR), transcription-mediated amplification (TMA) and branched DNA assay, allow the amount of HCV RNA in the blood, known as the viral load, to be estimated, which is useful in predicting response to treatment (see also Section 3.2.4) (Strader et al. 2004). Despite the high sensitivity of all these methods (Colin et al. 2001), one shortcoming is that they only assess the presence of HCV for diagnosis, and do not give an assessment of the severity of disease progression (Pawlotsky 2002). For prognostic assessment, a liver biopsy is most useful and is considered the 'gold standard'. However, this also has shortcomings, as it relies on medical expertise for correct interpretation, has a higher cost and is not without risk (Cadranel et al. 2000; Ghany et al. 2009).

3.2.4 Treatment

The current standard of care treatment for HCV is combined pegylated interferon alpha and ribavirin, with treatment lasting from 24 weeks for genotypes 2 and 3 to 48 weeks for genotype 1 and 4 (Fusco and Chung 2012). The treatment is associated with severe side effects, including fever, muscle pains, anorexia, hair loss, anaemia, thyroid problems and depression, and thus patients often choose or are forced to discontinue

with medication. Overall response rate in those who do proceed with treatment is 33-82%, but rates vary according to dose, genotype, gender, age when infection occurred and race (Manns et al. 2001; Fried et al. 2002). Early response to treatment, with a steep decline in HCV RNA by four weeks into treatment, is associated with a positive outcome (Ballesteros et al. 2004; Napoli et al. 2005), and patients are said to have cleared infection if they continue to have undetectable levels of HCV RNA 24 weeks after finishing treatment. However, those patients who fail to see greater than a 100-fold reduction in HCV RNA by week 12 are known as null responders, and treatment is discontinued due to the low chance of cure (Ghany et al. 2009). There are two other possibilities: partial responders show good decline until week 12, but do not clear HCV RNA by the end of the treatment period, and relapsers are those who had cleared infection by the end of treatment but show evidence of HCV RNA again within the 24 week follow-up period (Fusco and Chung 2012).

There are three additional treatments that can be used alongside interferon and/or ribavirin therapy to increase the chances of viral clearance. Boceprevir and telaprevir are protease inhibitors, which selectively bind to the viral enzymes involved in the breaking down of proteins, therefore blocking viral production; boceprevir targets the NS3 serine protease, and telaprevir targets the NS3/4A protease (Lin et al. 2006; Bogen et al. 2009). Both inhibitors are specific to HCV genotype 1, and have increased the response in chronic HCV genotype 1 infection (in both previously treated and untreated individuals), compared to standard interferon and ribavirin therapy alone (McHutchison et al. 2009; McHutchison et al. 2010; Bacon et al. 2011; Jacobson et al. 2011;

Poordad et al. 2011). Sofobuvir works slightly differently, metabolising to a substrate that inhibits the activity of the RNA polymerase in the NS5B protein, therefore restricting replication (Sofia et al. 2010). It can be used with ribavirin for interferon-free treatment of genotypes 1, 2, and 3, and can also be used alongside pegylated interferon to treat genotype 4 (Gilead 2013). Two of the major advantages of interferon-free sofobuvir treatment is that side effects tend to be less adverse and the dosage is taken orally rather than by subcutaneous injection, helping patient treatment adherence (Gane et al. 2013; Lawitz et al. 2013).

3.3 Modelling

3.3.1 Metapopulation Model for Pathogen Populations

Populations may be subdivided into a fixed number of subpopulations, or 'demes' (here, the terms are used interchangeably, following the notation of Wakeley as in, for example, Wakeley (1998)). Subpopulations may be completely or partially isolated from each other and the remainder of the population, perhaps physically or reproductively. Consequently the individuals within a subpopulation are seen to be more alike than the rest of the population, and this needs to be accounted for.

A metapopulation is classically defined as a 'population of populations' (Levins 1968a; Levins 1968b), and is taken here as a population that can be subdivided into a number of demes. Previously the notion of a metapopulation has been used to describe heterogeneity in pathogen species caused by host or strain structure (May and Nowak 1994; Bahl et al. 2011). However, the idea can be used in a more fundamental sense, as

members of a pathogen population can be split into subpopulations according to the individual they infect. In this scenario, demes then refer to infected individuals or hosts, and the total pathogen population is the aggregate of these demes. The main advantage of thinking about a pathogen population as a metapopulation, as opposed to merely a structured population, is that it allows for a deme to go extinct, and subsequently be re-colonised by other demes (Slatkin 1977). Under this framework, transmission is thus equivalent to an organism migrating from one host to another. Primary infection occurs when an empty deme (host) is colonised, and secondary infection when the deme has already been colonised (Figure 3.3.1). Recovery occurs when the pathogen population is cleared, and thus when a deme goes extinct.

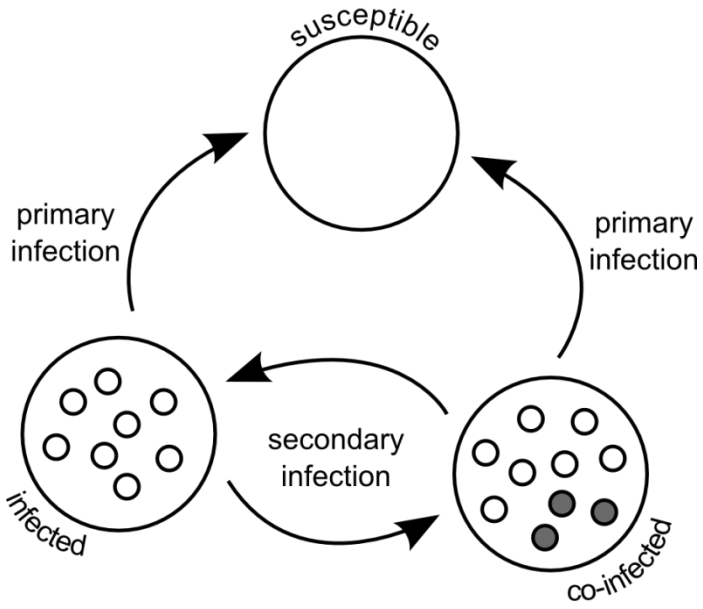


Figure 3.3.1: Metapopulation dynamics for infectious disease. Pathogen populations are split into subpopulations according to the individual they infect. Colonisation of a susceptible host is known as primary infection, and any further transmission events are referred to as secondary infection.

One of the advantages of using a metapopulation model is the wealth of understanding already available in the literature (Wright 1940; Levins 1968a; Levins 1968b; Slatkin 1977; Pannell and Charlesworth 1999). Wakeley derived a number of coalescent approximations for structured populations (Wakeley 1998; Wakeley 1999; Wakeley 2001), and the main result for his work on metapopulations (Wakeley and Aliacar 2001; Wakeley 2004b) is summarised in the next section.

3.3.2 Metapopulation Coalescent

For large numbers of demes (interpreted as having a sample size much smaller than the number of demes), the genealogy can be split into two main stages that Wakeley and Aliacar (2001) denote the scattering phase and the collecting phase. During the scattering phase, demes containing more than one lineage are reduced to a single lineage backwards in time through a Wright-Fisher process. This is due to a combination of coalescence within each deme, and extinction and migration events to other subpopulations that do not already contain lineages ancestral to the sample. At the end of this phase, ready for the start of the collecting phase, each different lineage that remains is in a separate deme.

Since each deme only contains a single lineage, during the collection phase migration and extinction events are essentially the same – simply moving the lineage to a different deme. When the number of demes, D, is large, most of the time is spent waiting for two separate lineages to enter the same deme. Since migration occurs with much higher frequency than coalescence, the probability of this occurring in the same deme is in the

order of $1/D$ and therefore rare. Once two lineages have entered the same deme, they can coalesce or migrate. When coalescence is rare the majority of migration events do not result in a coalescent event, and therefore on the whole time scale the migration events relevant to the genealogy lead to instantaneous coalescence. Thus, only migration events linked to coalescence events need to be considered explicitly, and the collecting phase to the time of the most recent common ancestor (MRCA) converges to a standard coalescent with a change in time scale. Moreover, since the collecting phase relies on rarer events than the scattering phase, it lasts for a much higher proportion of the history. Indeed, as the number of demes tends to infinity, the length of the scattering phase becomes negligible compared to that of the collecting phase.

So far the metapopulation coalescent has been thought about backwards in time. Since it will be used later in this chapter to model infection, it makes more sense to consider parameters forwards in time. Then, extinction is equivalent to colonisation (primary infection) forwards in time, and migration is equivalent to secondary infection. The many-demes limit (Wakeley and Aliacar 2001; Wakeley 2004b) gives the effective population size as:

$$N_e = \frac{D}{2(e_0 + m)F}, \tag{3.1}$$

where

$$F = \frac{1 + e_0 N_P/k}{1 + e_0 N_P/k + 2mN_P}.$$

In these equations, D is the number of demes as mentioned previously, e_0 is the rate of colonisation, m is the rate of migration, N_P is the pathogen population size within a host, and k is the number of genotypes in the founding population. F is known as the

inbreeding coefficient, and represents the probability of two lineages sampled from the same host are descended from the same transmission event (or equivalently, coalesced during the scattering phase).

3.3.3 Epidemiological Models

Compartmental models are important tools for modelling infectious disease dynamics, and have been widely applied in the understanding of infectious disease dynamics in host populations (Hethcote 2000). In the simplest case, the SI model, the population is split into the proportion of hosts who are susceptible, S, and those who are infectious, I. It is natural to expect that the primary rate of infection is dependent on the current density of infectious individuals and a transmission coefficient (β_1), known as strong proportionate mixing (Anderson and May 1991). This transmission coefficient combines all of the external factors that may affect transmission, such as environmental and social factors, and is unaffected by changes in epidemiology. In the SIS model, individuals recover and return to the susceptible group with the clearance rate γ, whereas in the SIR model, recovered hosts become immune from re-infection (Kermack and McKendrick 1927). The average length of infection is given by the inverse of the clearance rate, $1/\gamma$. All of these scenarios are illustrated in Figure 3.3.2.

One addition to these standard compartmental models, with a view to the metapopulation model, is the ability for already infected individuals to gain further infections, known as mixed, co- or secondary infection. Under strong proportionate mixing, secondary infection is dependent on the square of the number of infectious

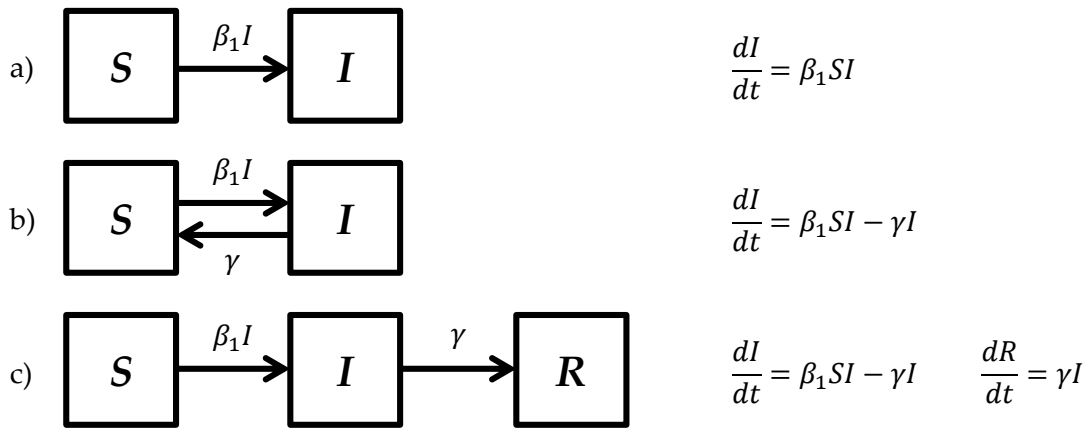


Figure 3.3.2: Compartmental models. The a) SI, b) SIS and c) SIR models can be modelled using differential equations for the change in the proportion of individuals susceptible (S), infected (I) or recovered (R). In all three models, the proportion of infected hosts increases at rate $\beta_1 SI$, where β_1 is the primary transmission coefficient. In the SIS model, individuals clear infection and return to the susceptible group. In the SIR model, individuals instead recover and are no longer susceptible. Note that dependency on time is suppressed for clarity of notation.

individuals and a transmission co-efficient, β_2. This secondary infection does not affect the overall prevalence given by I, as the individual is already included in the infected population.

Figure 3.3.3 illustrates the dynamics of the SI, SIS and SIR models. At the start of an epidemic, the prevalence is low and there is a high availability of susceptible hosts. The proportion of infected hosts then rapidly increases, growing exponentially with the intrinsic growth rate, r_0. In the SI model, the intrinsic growth rate is equal to β_1, and in the SIS and SIR models, $r_0 = \beta_1 - \gamma$. However, as the susceptible population gets exhausted, the prevalence increases at a slower rate. In the SI and SIR models, this continues until no susceptible individuals remain. In the SIS model, the infection continues to persist in dynamic equilibrium, where the rate that susceptible individuals become infected is balanced by those who clear infection and return to the susceptible group. In the SIR model, once the susceptible population has been depleted the

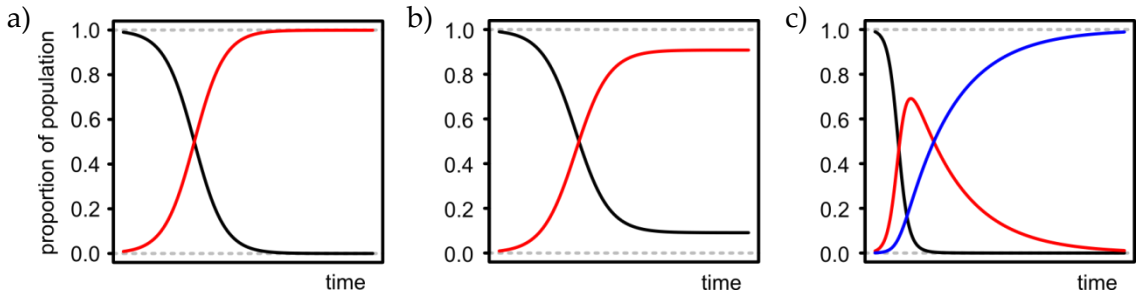


Figure 3.3.3: Epidemiological dynamics of compartmental models. The dynamics for the proportion of individuals susceptible (black), infected (red) and recovered (blue) change over time. In the SI model (a) eventually all of the population becomes infected. In SIS model (b), dynamic equilibrium is attained, and in the SIR model (c), the infection dies out and eventually the whole population becomes immune.

epidemic peaks and burns out as individuals recover. Modelling prevalence with deterministic differential equations in this way assumes that the number of infected hosts is large. This is clearly an approximation because there cannot be a large number of infected hosts at the start of an epidemic. However, previous work suggests that such approximations are still useful for epidemiological inference (Ferguson et al. 2001; Lipsitch et al. 2003; Fraser et al. 2009).

As previously mentioned, one important parameter in quantifying the dynamics of infectious disease is the basic reproductive number, R_0. As well as comparing the transmissibility of disease, it can also be used to determine the effect that treatment or vaccination intervention is having (Anderson and May 1991). When R_0 is less than one, there is not enough transmission between individuals to sustain the infection, and the disease will eventually go extinct. Hence an infection can only persist in the host population if R_0 is greater than or equal to one. The basic reproductive number is undefined in the SI model, and is given by $R_0 = \beta_1/\gamma$ in the SIS and SIR models.

3.3.4 Combined Epidemiological and Coalescent Inference

The previous section introduced compartmental modelling for epidemiology; these parameters are now used to specify the coalescent metapopulation model from Section 3.3.2. For reference, all parameter definitions used in the Modelling section of this chapter can be found in Table 3.3.1.

Letting the total number of infected hosts be denoted N_H, the number of occupied (infected) demes, D, is given by $N_H I$. The rates of primary transmission (colonisation) and secondary transmission (migration) are $e_0 = \beta_1 S$ and $m = \beta_2 I$ respectively. Then the equation for the effective population size of the pathogen population given in Equation (3.1) becomes

$$N_e = \frac{N_H I}{2(\beta_1 S + \beta_2 I)F}, \tag{3.2}$$

where

$$F = \frac{N_P^{-1} + \beta_1 S/k}{N_P^{-1} + \beta_1 S/k + 2\beta_2 I}.$$

The effective population size is given by a complex function of several epidemiological parameters. In particular, the inbreeding coefficient, F, and the primary and secondary rates of infection all depend on the prevalence. This dependency on the prevalence resolves the conflicting observations about the proportionality of prevalence and the effective population size at equilibrium (Koelle and Rasmussen 2012), but increases in prevalence do not linearly increase the effective population size (Frost and Volz 2010).

Table 3.3.1: Summary of parameters used in combined metapopulation and epidemiological coalescent inference.

Parameter	Description
Epidemiological variables	
S	Proportion of hosts that are susceptible
I	Proportion of hosts that are infected
R	Proportion of hosts that are recovered
Epidemiological parameters	
N_H	Total number of hosts
S_0	Proportion of hosts susceptible in the present
β_1	Primary transmission coefficient
β_2	Secondary transmission coefficient
γ	Rate of clearance of infection
$R_0 = \beta_1/\gamma$	Basic reproductive number (SIS and SIR models)
Metapopulation parameters	
N_e	Effective population size of the metapopulation
F	Inbreeding coefficient
$D = N_H I$	Total number of infected hosts
$e_0 = \beta_1 S$	Rate of primary transmission per infected host
$m = \beta_2 I$	Rate of secondary transmission per infected host
N_P	Pathogen population size within an infected host
k	Number of haploid pathogens transmitted during primary infection
SI model: coalescent parameters*	
$N_0 = \frac{N_H(1-S_0)}{2\beta_1 S_0}$	Effective population size of the metapopulation at the present
$r_0 = \beta_1$	Intrinsic growth rate of the epidemic
SIS model: coalescent parameters*	
$N_0 = \frac{N_H(1-S_0)}{2\beta_1 S_0}$	Effective population size of the metapopulation at the present
$r_0 = \beta_1 - \gamma$	Intrinsic growth rate of the epidemic
$t_{50} = -\frac{\log\left(\frac{\beta_1-\gamma}{\gamma(1-S_0)}-1\right)}{\beta_1-\gamma}$	Time at which N_e reached half its maximum
SIR model: coalescent parameters*	
$N_0 = \frac{N_H\left(1-S_0+\frac{\gamma}{\beta_1}\log(S_0)\right)}{2\beta_1 S_0}$	Effective population size of the metapopulation at the present
$r_0 = \beta_1 - \gamma$	Intrinsic growth rate of the epidemic
$t_{peak} = S^{-1}\left(\frac{\gamma}{\beta_1}\right)$	Time of peak prevalence. $S^{-1}(s)$ represents the time at which the proportion of susceptibles equals s, which must be computed numerically.

* Assuming co-infection is negligible (i.e. $\beta_2 = 0$).

It is worth noting that when co-infection is rare, there is little within-host diversity. This means that recombination will have little to no effect on either the genealogy or genetic diversity. However, when secondary infection is present, the number of mixed infections will increase as the epidemic grows in size. This causes the effective recombination rate to increase, as more lineages will have the ability to come into contact and thus exchange genetic information. For the rest of this chapter, it is assumed that co-infection is negligible ($\beta_2 = 0$), an assumption that is both reasonable in many cases of interest, and substantially reduces the difficulty in estimating from real data. In the absence of recombination, all sequences sampled from the same individual must have descended from the founding ancestor when infection took place, and therefore the inbreeding coefficient F equals one.

3.3.5 Coalescent SI and SIS Models

The SI and SIS models can be solved in closed form through integrating the equations given in Figure 3.3.2. Note that by setting the present time (the time of sampling) at $t = 0$, and working backwards in time with $t > 0$, the rate of change must be negative to ensure the epidemic decreases in size back towards the origin. For the SI model,

$$S = \frac{S_0}{S_0 + (1 - S_0)e^{-\beta_1 t}}$$
$$I = 1 - S, \tag{3.3}$$

and for the SIS model,

$$S = \frac{\beta_1 S_0 - \gamma + \gamma(1 - S_0)e^{-(\beta_1 - \gamma)t}}{\beta_1 S_0 - \gamma + \beta_1(1 - S_0)e^{-(\beta_1 - \gamma)t}}$$
$$I = 1 - S. \tag{3.4}$$

The parameter S_0 is interpreted as the proportion of individuals susceptible in the present. These equations can then be put in to Equation (3.2) to obtain a simplified form for the effective population size under the metapopulation model. For the SI model, the effective population size simplifies to:

$$N_e = N_0 e^{-r_0 t}, \tag{3.5}$$

where

$$r_0 = \beta_1$$

$$N_0 = \frac{N_H(1 - S_0)}{2\beta_1 S_0}.$$

This is an exponential growth curve, where r_0 is the intrinsic growth rate, and N_0 is the effective population size at present. For the SIS model, the effective population size becomes:

$$N_e = N_0 \frac{1 + e^{-r_0 t}}{1 + e^{r_0(t_{50} - t)}}, \tag{3.6}$$

where

$$r_0 = \beta_1 - \gamma$$

$$N_0 = \frac{N_H(1 - S_0)}{2\beta_1 S_0}$$

$$t_{50} = \frac{-\log\left(\frac{r_0}{\gamma(1 - S_0)} - 1\right)}{r_0}.$$

This is a logistic growth curve with parameters r_0 and N_0 defined as before, and t_{50}, the time at which the effective population sized has reached half its carrying capacity.

Equations (3.5)and (3.6) show that under standard mixing assumption and negligible secondary transmission, the SI and SIS models resemble two standard coalescent demographic models widely used in analyses of pathogen dynamics (Pybus et al. 2000;

Nakano et al. 2004; Pybus et al. 2005). However, the growth curves for the effective population size are simpler and have one fewer parameter than those describing changes in the prevalence over time. As a result, there is not a one-to-one correspondence between the coalescent and epidemiological parameters. Importantly, this means that an independent estimate of one of the epidemiological parameters is required to separate the others to, for example, reconstruct the historical prevalence of infection. This differs from the results of Pybus et al. (2001), who do not require an independent estimate to recover the trajectory of prevalence; however, the method outlined here agrees that a) the intrinsic growth rate r_0 in an SIS model can be estimated by a logistic growth curve for the effective population size, and b) that an independent estimate of an epidemiological parameter is required to obtain a value of R_0.

3.3.6 Coalescent SIR Model

Unlike in the SI and SIS models, the SIR model cannot be solved analytically for S. However, the system of differential equations given in Figure 3.3.2 can be simplified to a single ordinary differential equation by assuming that the number of recovered (or immune) individuals is zero at the start of the epidemic, implying the relationship:

$$I = 1 - S + \frac{\gamma \log(S)}{\beta_1}. \tag{3.7}$$

Then, the following equation can be solved numerically:

$$\frac{dS}{dt} = \beta_1 S(1 - S) + \gamma S \log(S)\,. \tag{3.8}$$

As explained in Section 2.5.3, when considering demographic growth in the coalescent, the pairwise coalescent rate is equal to the inverse of the effective population size, and

the cumulative rate of coalescence is given by $\Lambda(t) = \int_0^t 1/N_e(u)du$. Assuming no co-infection, this integral can be written as a differential equation:

$$\frac{d\Lambda}{dt} = \frac{1}{N_e} = \frac{\left(1 - S_0 + \frac{\gamma \log(S_0)}{\beta_1}\right) S}{N_0 S_0 \left(1 - S + \frac{\gamma \log(S)}{\beta_1}\right)}, \tag{3.9}$$

where

$$N_0 = \frac{N_H \left(1 - S_0 + \frac{\gamma \log(S_0)}{\beta_1}\right)}{2\beta_1 S_0}.$$

Since the effective population size is dependent on S, Equations (3.8) and (3.9) define a system of differential equations that can be solved together. Importantly, unlike for the SI and SIS models, there is no confounding – meaning that the epidemiological parameters including R_0 can be estimated from the genetic analysis alone.

3.4 Methods

3.4.1 Genetic Data

I performed a literature search for hepatitis C datasets with a well-described sampling frame, and for which information about subtype, sampling location, prevalence and *NS5B* gene sequences were available online. From an initial search, 28 datasets were identified (Table 3.4.1); this was further narrowed to 18 due to the exclusion of any with a sample size fewer than 20 sequences (NakBra1a, NakUS1b and AkkTha6n), evidence of recombination (NakUS1a, NakVie1a, NakBra1b, NakChiA1b, TanSoA5a and FuChi6a), or were found to have questionable sampling on more thorough investigation (NakChiB1b). Genbank accession numbers for the sequences in all these datasets are given in Appendix A.

Table 3.4.1: Summary of hepatitis C datasets collated after literature review.

Dataset	Sub Type	Country	Study Period	N	Recombination tests r^2	|D'|	G4	Source
MatBel1a	1a	Belgium	2009	45	0.971	0.854	0.841	a
NakBra1a	1a	Brazil	2009	18	0.721	0.875	0.862	b
NakInd1a	1a	Indonesia	2004-2007	30	0.793	0.644	0.630	b
NakUS1a	1a	US	2004-2007	24	0.211	0.658	0.001	b
NakVie1a	1a	Vietnam	1991-2002	27	0.036	0.215	0.229	b
PhaVie1a	1a	Vietnam	2005	21	0.42	0.584	0.723	c
PybUK1a	1a	UK	2000	52	0.509	0.255	0.188	d
TanUS1a	1a	US	2000	30	0.81	0.849	0.868	e
FuChi1b	1b	China	1999-2002	120	0.447	0.195	0.133	f
KurMon1b	1b	Mongolia	1999-2002	60	0.647	0.466	0.457	g
NakBra1b	1b	Brazil	2002	26	0.001	0.57	0.627	b
NakChiA1b	1b	China	2002	35	0.039	0.69	0.692	h
NakChiB1b	1b	China	1999-2002	30	0.304	0.165	0.253	h
NakInd1b	1b	Indonesia	1999-2002	37	0.103	0.935	0.925	b
NakUS1b	1b	US	1999-2002	17	0.011	0.771	0.727	b
NakVie1b	1b	Vietnam	1999-2002	32	0.594	0.714	0.712	b
TanSpa1b	1b	Spain	1999-2002	33	0.273	0.166	0.136	e
ReArg2c	2c	Argentina	1999-2002	75	0.538	0.093	0.112	i
MatBel3a	3a	Belgium	2008	47	0.115	0.705	0.746	a
PybUK3a	3a	UK	2008	63	0.368	0.087	0.097	d
TanEgy4a	4a	Egypt	1997-2001	47	0.867	0.957	0.954	e
HenFra5a	5a	France	1997-2001	150	0.067	0.762	0.799	j
TanSoA5a	5a	South Africa	1999-2008	24	0.044	0.916	0.917	e
FuChi6a	6a	China	2000-2003	77	0.004	0.28	0.405	f
PhaVie6a	6a	Vietnam	2000-2003	24	0.533	0.917	0.925	c
TanHon6a	6a	Hong Kong	2000-2003	23	0.435	0.631	0.638	e
AkkTha6f	6f	Thailand	2000-2003	39	0.073	0.638	0.552	k
AkkTha6n	6n	Thailand	2000-2003	16	0.382	0.176	0.162	k

Abbreviations: ***n*** Number of sequences. Source: **a** Matheï et al. (2008). **b** Nakano et al. (2004). **c** Pham et al. (2009). **d** Pybus et al. (2005). **e** Tanaka et al. (2006). **f** Fu et al. (2011). **g** Kurbanov et al. (2007). **h** Nakano et al. (2006). **i** Ré et al. (2011). **j** Henquell et al. (2011). **h** Akkarathamrongsin et al. (2010). Red highlights the reason for exclusion from final analysis.

Recombination contradicts the modelling assumptions and is problematic in phylogenetic analyses when not explicitly counted for as the resulting phylogeny looks star-like – a feature that is also seen under exponential growth (Schierup and Hein 2000). This is clearly a problem when trying to assess growth rates. Recombination also provides evidence of co-infection, which is assumed to be negligible (Section 3.3.4). To assess evidence of recombination in each dataset, a simple permutation test was used. Permutations of the sequences are made by randomly reordering the nucleotide positions, and the observed correlation between physical distance in the genome and three measures of linkage disequilibrium (r^2, $|D'|$ and G4) is calculated as a part of omegaMap (Wilson and McVean 2006). Under the null hypothesis of no recombination, all nucleotide sites must be equally likely to be linked, regardless of how far apart they are located in the genome. A dataset was excluded from the final analysis if the null hypothesis was rejected at the 5% level by any of the three tests (highlighted in red in Table 3.4.1). The Geneious v.5.6 (Kearse et al. 2012) alignment tool was used to produce a global alignment of sequences across all of the datasets, and, where an alignment was not available, for sequences in the same dataset. These alignments are available online in Dataset S1 of the supplementary material of (Dearlove and Wilson 2013).

3.4.2 Model Averaging Approach

The Bayes factor (see Section 2.4.1) is a standard method of model selection in the Bayesian setting. Whilst the simplest method of estimating the marginal likelihood is the harmonic mean estimator (HME), only requiring the posterior samples from an MCMC run (Newton and Raftery 1994), it can be biased by samples with small likelihoods

leading to an overestimate of the true marginal likelihood (Xie et al. 2011). Since other methods such as stepping-stone sampling and path sampling (Baele et al. 2012) were not available in BEAST at the time of analysis, a mixture model was implemented. Letting X be the observed data, M the model, and θ the union of all model parameters including the genealogy, the posterior is given by:

$$\begin{aligned} p(\theta|X) &= \sum_M p(\theta, M|X) \\ &= \sum_M \frac{p(X|\theta, M)p(\theta|M)p(M)}{p(X)}, \end{aligned} \tag{3.10}$$

where $p(X|\theta, M)$is the likelihood of the data conditional on the model parameters θ given by Felsenstein's pruning algorithm. The marginal likelihood, which gives the evidence towards the model in question, is

$$p(X|M) = \int_{\theta\in\Theta} p(X|\theta, M)p(\theta|M)d\theta, \tag{3.11}$$

which can be estimated using importance sampling with a proposal distribution, $q(\theta)$:

$$\begin{aligned} p(X|M) &= \int_{\theta\in\Theta} p(X|\theta, M)\frac{p(\theta|M)}{q(\theta)}\, q(\theta)d\theta \\ &\approx \frac{1}{n}\sum_{i=1}^{n} p\left(X\middle|\theta^{(i)}, M\right)\frac{p\left(\theta^{(i)}\middle|M\right)}{q(\theta^{(i)})}, \qquad \theta^{(i)}\sim q(\theta). \end{aligned} \tag{3.12}$$

Setting the proposal distribution as $p(\theta|X)$ in Equation (3.10), yields

$$\begin{aligned} p(X|M) &\approx \frac{1}{n}\sum_{i=1}^{n} \frac{p\left(X\middle|\theta^{(i)}, M\right)p\left(\theta^{(i)}\middle|M\right)}{\sum_{m'} p\left(X\middle|\theta^{(i)}, M\right)p\left(\theta^{(i)}\middle|M\right)p(M)\,/p(X)}, \\ &\theta^{(i)}\sim p(\theta|X). \end{aligned} \tag{3.13}$$

Therefore, it can be shown that an importance sampling estimate of the posterior probability of model m can be calculated as:

$$p(M=m|X)=\frac{p(M=m)p(X|M)}{p(X)}$$
$$\approx\frac{1}{n}\sum_{i=1}^{n}\frac{p\left(X\middle|\theta^{(i)},M=m\right)p\left(\theta^{(i)}\middle|M=m\right)p(M=m)}{\sum_{m'}p\left(X\middle|\theta^{(i)},M=m'\right)p\left(\theta^{(i)}\middle|M=m'\right)p(M=m')}, \tag{3.14}$$

where

$$\theta^{(i)}\sim p(\theta|X).$$

This equation can be further simplified, since $p(X|\theta,M=m)=p(X|\theta)$. Using this method, many models can be compared in a single MCMC run, and there is no requirement that they need be nested (for example, the SI model is nested in both the SIS and SIR models by setting $\gamma=0$, but neither the SIS or SIR can be simplified in such a way that gives the other). When parameters are shared between models, it is also possible to obtain estimates of parameters averaged over all of the possible models. When a model has a high posterior probability, this will dominate the overall parameter estimate and the model averaging will not have a huge effect, whereas in a situation with less conclusive evidence towards the best model the averaging may hint towards why a clear distinction cannot be made.

3.4.3 BEAST Analysis

Five different analyses were run for each dataset – the endemic (implying a constant effective population size), SI, SIS and SIR models, and one under the model averaging approach given in Section 3.4.2. The SIR model was implemented as an extension to BEAST in Java by Daniel Wilson, using a fifth-order Cash-Karp Runge-Kutta method with adaptive stepsize control (Press et al. 2002). A re-parameterisation of the default logistic growth model was also implemented by Daniel Wilson, so that N_0 gives the effective population size at the present, as opposed to the default which gives the

carrying capacity. The model averaging approach uses functions already implemented in BEAST, but requires some extra coding in the input XML file. An outline of how to edit the standard demographic model for this purpose is given in Appendix B.

The analysis was performed in BEAST v1.7 (Drummond et al. 2012). An improper log-uniform prior was assumed for N_e, an exponential prior with mean one year for r_0, an exponential prior with mean one year for γ and an exponential prior with mean 50 years for t_{50} and t_{peak}. The HKY85 model of nucleotide substitution (Hasegawa et al. 1985) was used with a fixed mutation rate of 5.8×10^{-4} substitutions per site per year (Tanaka et al. 2002) for calibration purposes, since not all of the datasets had information regarding sampling dates. This rate was estimated for the *NS5B* gene, and was deemed reasonable given that it had previously been applied to a number of the datasets covered by this analysis. A log-normal prior with a mean of 1.0 and standard deviation of 1.25 on the logarithmic-scale was assumed for the transition:tranversion ratio, κ, with a uniform distribution on the nucleotide frequencies. Equal prior probabilities for each model were assumed for the model averaging approach.

Two independent Markov Chain Monte Carlo (MCMC) chains were run for each analysis. For the model averaging approach and SIR model, the MCMC was run for a chain of 100 million steps, with samples taken every 1000 iterations and the first 10% of samples discarded as burn in. All other models were run for 50 million iterations. Results quoted use the median for point estimates, and the 2.5% and 97.5% quantiles for credible intervals. For the meta-analysis, the effective population size at the time of

sampling (N_0), intrinsic growth rate (r_0) and time to most recent common ancestor were estimated using the model averaging approach. The doubling time was calculated as ($\log(2)/r_0$).

3.4.4 Investigating Hepatitis C Diversity

The genetic diversity for each HCV dataset was measured using π, the mean number of pairwise differences between sequences in the same dataset. To identify which epidemiological processes have a significant association with diversity, a linear regression was used. Demographic covariates of interest included the host population size and density, reported prevalence and HCV subtype. Estimates for the host data and prevalence were taken to be those of the country in which the data was sampled (unless in the case of prevalence there was a clear local discrepancy), and where possible, were obtained from the same sampling time frame as the sequence data. Whilst prevalence in terms of the model is taken to be the proportion of the population who are infected at any point in time, the reported prevalence here (given in Table 3.5.1) includes studies of seropositivity as well as disease. The seroprevalence is the proportion of people in which antibodies to HCV are detected, and thus is subtly different to the prevalence, as it represents everyone who has been infected in their lifetime, rather than just those infectious at the current time (Gelberg et al. 2012). However, such studies represent a 'best guess' for the true disease incidence in the population, which is often otherwise infeasible to estimate (Shepard et al. 2005).

The median of the model averaged posterior distributions of the time to most recent common ancestor, T_{MRCA}, intrinsic growth rate, r_0, and effective population size, N_0, obtained in the BEAST analysis were also included in the regression. A stepwise regression approach using the *F*-test criterion was taken to decide the explanatory variables that should remain in the model in addition to r_0, which was included due to a strong prior interest in the effect of the intrinsic growth rate.

3.5 Results

3.5.1 Diversity of Hepatitis C Epidemics

The geographical location and subtypes of the final 18 datasets are summarised in Figure 3.5.1a, with colours differentiating between datasets of the same subtype. The maximum likelihood genealogy of the global alignment in Figure 3.5.1b shows that subtypes form distinct monophyletic groups, which would be expected given the high diversity between groups (see overview in Section 3.1). However, this is not always the

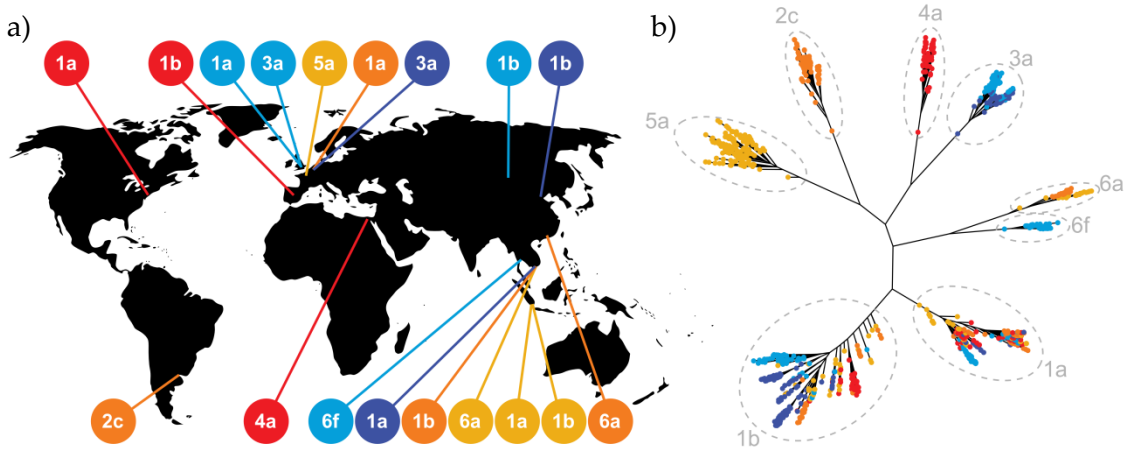


Figure 3.5.1: Summary of hepatitis C datasets. a) The geographical location of the HCV datasets analysed, annotated by subtype. Colours are used to differentiate between datasets of the same subtype. b) A maximum likelihood genealogy of all sequences based on a global alignment of the NS5B gene. Colours are as in a). To aid with visualisation, a square root transformation was applied to branch lengths. Pre-print: Dearlove and Wilson (2013).

case for the ancestral histories of datasets within subtypes. In subtype 6a, the two datasets form distinct clusters of a single colour, whereas in subtypes 1a and 1b there is much more overlap in relatedness between datasets.

Table 3.5.1 shows the collated demographic data for the linear regression, including the genetic derived data and the host population data. The average pairwise nucleotide diversity was found to vary widely across the 18 datasets, ranging from 38.2 nucleotide differences per kilobase for subtype 3a in Belgium to 84.3 for subtype 4a in Egypt. The model averaged T_{MRCA} is seen to be less than 100 years in all but one case (subtype 4a in Egypt).

The results of fitting the regression model are given in Table 3.5.2 and plotted in Figure 3.5.2. Note that the quoted p-values should be taken with caution, due to the overlap in ancestry within some subtypes. The strongest predictor of diversity was the time to the most recent common ancestor, T_{MRCA}. For each 10 years in the age to the MRCA, diversity increases by nearly 4.6 nucleotide differences per kilobase. In contrast, population density had a negative relationship with diversity, with an increase in population density for a dataset causing a decrease in the diversity; for every extra 100 people per square kilometre the diversity is expected to decrease by 2.87 nucleotides per kilobase. In the subtypes where there was more than one observation, subtype 1b had the highest diversity with, on average, 7.33 nucleotides per kilobase more than subtype 6a, which had the lowest diversity. There was no significant relationship between diversity and the intrinsic growth rate, r_0, after adjusting for the other factors.

Table 3.5.1: Demographic data for investigating diversity in HCV.

Dataset	Sequence derived data				Population data		
	π (kb^{-1})	T_{MRCA} (years)	r_0 ($year^{-1}$)	N_0 (10^3)	Size (10^6)	Dens. (km^{-3})	Prev. (%)
MatBel1a	45.0	64.42 (54.06, 77.12)	0.09 (0.05, 0.16)	0.72 (0.35,1.59)	10.4	340.7	0.9
NakInd1a	47.5	52.03 (36.92, 74.31)	0.10 (0.05, 0.23)	0.68 (0.23, 2.48)	218.7	114.8	2.1
PhaVie1a	54.4	79.25 (63.99, 98.47)	0.06 (0.03, 0.12)	0.65 (0.23, 2.12)	81.5	245.7	6.1
PybUK1a	47.7	57.14 (49.21, 66.76)	0.15 (0.11, 0.19)	6.86 (2.64, 20.05)	59.7	243.8	1.0[a]
TanUS1a	49.5	58.19 (48.16, 70.90)	0.15 (0.10, 0.23)	12.40 (1.70, 126.41)	293.6	30.5	1.8
FuChi1b	63.9	82.66 (70.63, 97.58)	0.08 (0.06, 0.10)	3.47 (2.27, 5.56)	1300.1	135.8	3.0
KurMon1b	63.0	75.79 (66.88, 86.77)	0.15 (0.11, 0.20)	259.54 (24.15, 2677.83)	2.5	1.6	15.0[b]
NakInd1b	59.0	71.31 (58.65, 87.47)	0.11 (0.08, 0.17)	11.79 (2.15, 81.35)	218.7	114.8	2.1
NakVie1b	47.3	59.70 (43.41, 82.44)	0.08 (0.04, 0.17)	0.61 (0.23, 2.01)	81.5	245.7	6.1
TanSpa1b	62.4	76.53 (65.64, 90.69)	0.14 (0.09, 0.21)	86.65 (5.56, 2405.77)	42.5	84.0	2.5[b]
ReArg2c	63.1	81.18 (71.71, 92.88)	0.14 (0.10, 0.20)	5.04 (1.52, 17.48)	37.9	13.6	5.8[c]
MatBel3a	39.5	49.59 (42.48, 60.69)	0.29 (0.08, 0.67)	0.55 (0.35, 1.38)	10.4	340.7	0.9
PybUK3a	47.8	56.71 (50.62, 64.35)	0.17 (0.13, 0.24)	31.93 (4.28, 153.26)	59.7	243.8	1.0[a]
TanEgy4a	84.3	104.80 (93.25, 118.89)	0.11 (0.08, 0.15)	347.78 (21.88, 5199.30)	73.4	73.3	16.3[d]
HenFra5a	55.9	78.77 (69.73, 89.94)	0.12 (0.10, 0.15)	60.37 (16.53, 147.12)	60.0	108.8	1.3[b]
PhaVie6a	49.5	78.19 (58.60, 103.11)	0.08 (0.04, 0.15)	1.60 (0.51, 7.90)	81.5	245.7	6.1
TanHon6a	38.2	41.14 (33.30, 51.37)	0.37 (0.17, 1.22)	0.25 (0.06, 1.22)	6.8	159.2	0.5
AkkTha6f	41.8	46.64 (40.92, 53.92)	0.26 (0.18, 0.42)	14.63 (1.28, 699.01)	63.8	124.3	5.6

Abbreviations. π The average pairwise nucleotide diversity per kilobase. T_{MRCA} Time to most recent common ancestor. r_0 Intrinsic growth rate. N_0 Effective population size. **Dens.** Population density. Population size and density obtained from Population Reference Bureau (2004). **Prev.** Reported prevalence. Estimates of prevalence from World Health Organisation (1999), unless indicated: **a** Esteban et al. (2008). **b** Kurbanov et al. (2007)**. c** Mengarelli et al. (2006). **d** Pybus et al. (2003).

Table 3.5.2: Coefficients of the linear regression to understand HCV diversity.

Coefficient	Estimate	s.e.	*F*-test	*T*-test	*p*-value
Intercept	25.3	5.93	-	-	-
T_{MRCA}	0.456	0.0719	40.2	-	0.0004
Population density	-0.0287	0.00638	20.2	-	0.0028
Subtype:	-	-	6.48	-	0.0124
- 1b versus 1a	3.04	1.48	-	2.06	0.0782
- 2c versus 1a	0.222	2.73	-	0.081	0.9374
- 3a versus 1a	0.981	2.07	-	0.473	0.6507
- 4a versus 1a	12.6	3.68	-	3.41	0.0112
- 5a versus 1a	-3.05	2.43	-	-1.26	0.2495
- 6a versus 1a	-4.33	2.01	-	-2.15	0.0682
- 6f versus 1a	-2.85	2.34	-	-1.22	0.2631
r_0	6.373	11.7	0.297	0.297	0.6027

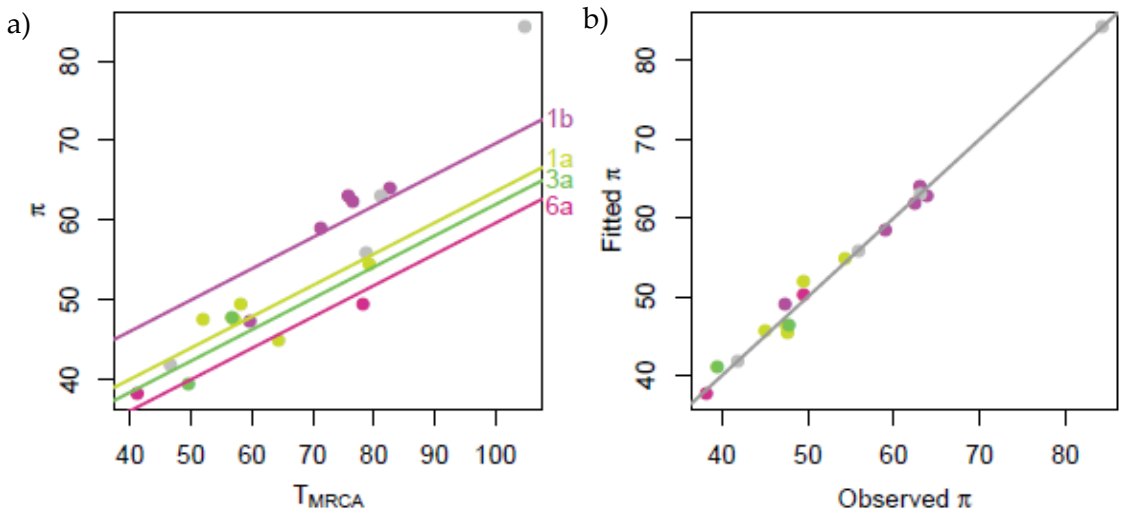


Figure 3.5.2: Meta-analysis of HCV diversity. Linear regression of nucleotide diversity (π) versus time to most recent common ancestor (T_{MRCA}), subtype, population density and intrinsic growth rate (r_0). a) Scatterplot of π against T_{MRCA}, with lines showing the effect of subtype where there were at least two observations. b) Plot of fitted diversity against observed diversity. Preprint: Dearlove and Wilson (2013).

3.5.2 Historical Effective Population Size

Reconstructing historical changes in N_e using the model averaging approach revealed that most datasets showed strong exponential growth corresponding to the SI model, as shown by the dominance of linear trajectories on the logarithmic scale in Figure 3.5.3. For each dataset, the posterior probability (PP) for each model (endemic, SI, SIS or SIR) was calculated, and the posterior probabilities for the latter three models are inset as a bar chart and also given in Table 3.5.3. The endemic model was rejected for all datasets (PP≤0.002) and thus is not shown in the plot. The preference for the SI model in 14 of the 18 datasets suggests that the dynamics in these epidemics have neither reached dynamic equilibrium (as in SIS) nor begun to burn out (as in SIR).

Only one example had the SIS model as most probable, Belgium subtype 3a (PP=87.9%), shown by the trajectory smoothing off from 1970 onwards. Interestingly, in the subtype 1a dataset from Belgium, SI dynamics had the highest posterior probability (PP = 44.0%), but there was also high support for the SIS (PP = 36.0%) and SIR (PP = 20.0%) models. Subtype 3a appears to be a much newer epidemic in Belgium that has reached dynamic equilibrium, whereas subtype 1a appears to be older and still growing. The high posterior probabilities for the SIS and SIR models might suggest, however, that the growth subtype 1a is on the verge of slowing down. Both of these observations are in stark contrast to the distinctly SI dynamics of the same subtypes in the UK, with posterior probabilities for the SI model equal to 92.5% and 61.6% in 1a and 3a respectively.

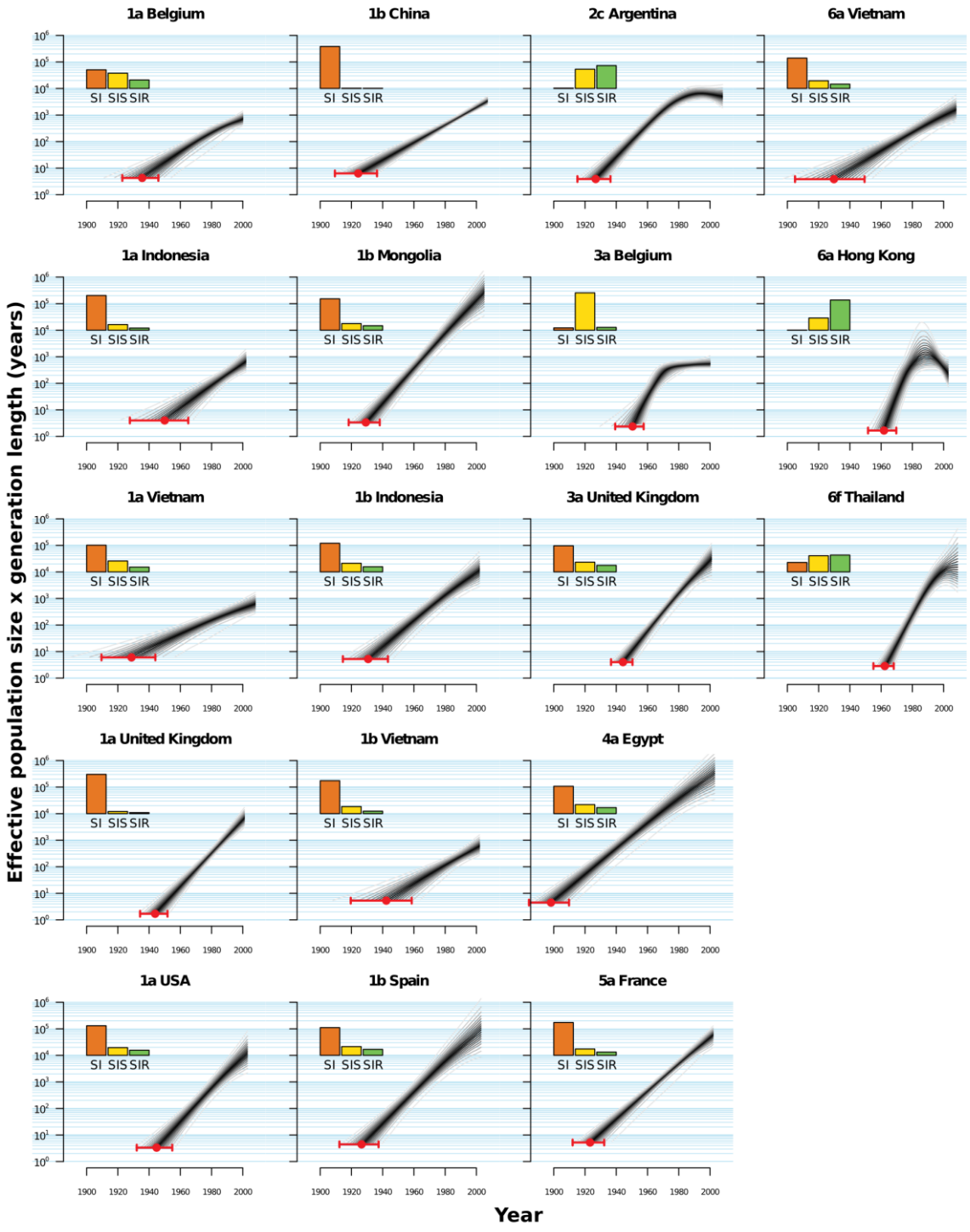


Figure 3.5.3: Reconstructed historical effective population size. For each of the 18 datasets, the reconstructed effective population size calculated by model average is plotted against time. Each dataset is labelled with the subtype and sampling location, and plotted against the same $\log_{10}$ scale. The grey lines show the quantiles of the posterior distribution of N_e averaged over all models (endemic, SI, SIS and SIR) at 5% increments. Quantiles closer to the median are darker. The red circle and error bar gives the time of the MRCA and associated 95% credible interval. Inset for each dataset is a bar plot giving the posterior probability of the SI, SIS and SIR models. The endemic model is not included at it was less than 0.2% for every dataset. Preprint: Dearlove and Wilson (2013).

Table 3.5.3: Posterior probabilities for the endemic, SI, SIS and SIR models.

		Posterior Probability (%)		
Dataset	**Endemic**	**SI**	**SIS**	**SIR**
MatBel1a	0.00	43.98	35.99	20.04
NakInd1a	0.01	81.70	13.29	5.00
PhaVie1a	0.23	62.75	25.71	11.30
PybUK1a	0.00	92.51	5.05	02.44
TanUS1a	0.00	69.71	18.21	12.09
FuChi1b	0.00	98.90	0.63	0.47
KurMon1b	0.00	73.92	15.50	10.58
NakInd1b	0.00	67.48	20.35	12.17
NakVie1b	0.07	77.56	16.61	5.77
TanSpa1b	0.00	65.301	20.60	14.09
ReArg2c	0.00	0.84	45.37	53.79
MatBel3a	0.00	5.44	87.93	6.64
PybUK3a	0.00	61.56	22.93	15.51
TanEgy4a	0.00	64.54	21.21	14.25
HenFra5a	0.00	77.37	14.97	7.66
PhaVie6a	0.00	71.59	18.17	10.24
TanHon6a	0.00	0.30	28.69	71.01
AkkTha6f	0.00	22.28	38.09	39.64

3.5.3 Examples of SIR Dynamics

In three cases, the SIR model was found to be the best – subtype 2c in Argentina, 6a in Hong Kong and 6f in Thailand. As mentioned in Section 3.3.6, in the SIR model the epidemiological parameters can be directly estimated from the genetic data. Therefore, for these three datasets it was also possible to estimate the historical prevalence of these epidemics, and to obtain an estimate for R_0. Due to the known relationship between N_0 and N_H in Equation (3.9), the total number of hosts infected throughout the epidemic can

be also be calculated and plotted alongside the prevalence (given as a proportion). These are both shown for all three epidemics with SIR dynamics in Figure 3.5.4.

The PP of the SIR model for subtype 2c in Argentina was 53.4%, with the SIS model the next likely with a PP of 45.4%. The T_{MRCA} was estimated to be between the years 1915 and 1936, with an initial doubling time of between 3.6 and 6.7 years (Table 3.5.4). The peak of the epidemic was estimated to have occurred between 1962 and 2002, and the prevalence has fallen since. HCV subtype 2c is generally uncommon; yet, in the Córdoba province of Argentina where this dataset was sampled, it is the most dominant - estimated to cause more than 50% of cases (Mengarelli et al. 2006; Ré et al. 2011). The

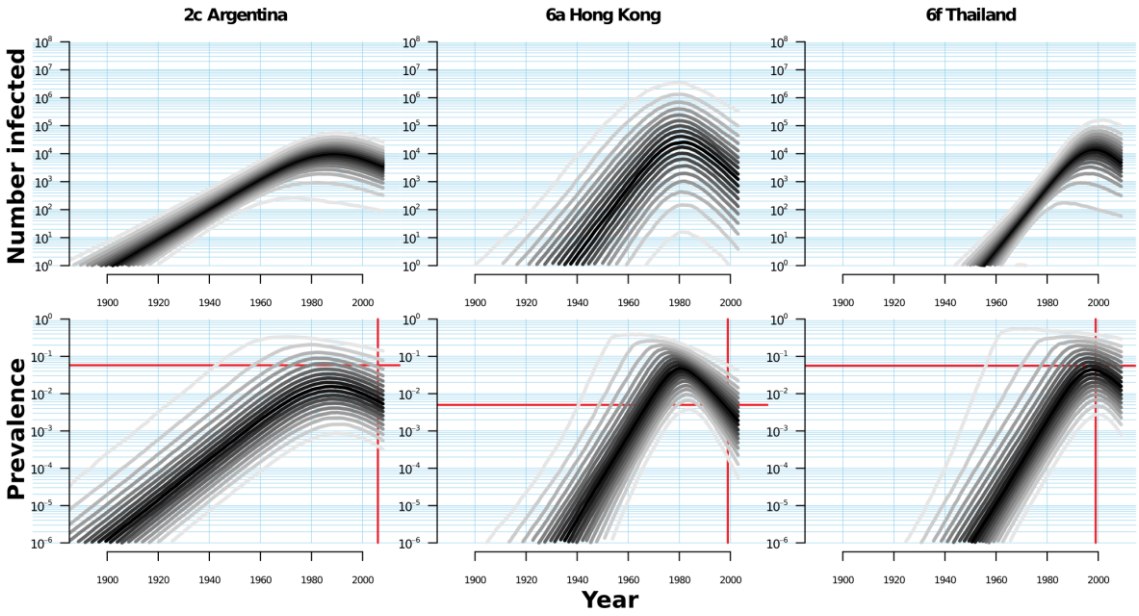


Figure 3.5.4: Reconstructed SIR dynamics for the number of infected hosts and prevalence. For subtypes 2c in Argentina, 6a in Hong Kong and 6f in Thailand, the model averaging approach showed preference for the SIR model. Under the SIR model, changes in the number of infected hosts and prevalence can be estimated directly from the genetic analysis. For each dataset, both of these are shown. The grey lines show the quantiles of the posterior distribution at 5% increments. Quantiles closer to the median are darker. The intersection of the red lines represents an independent point estimate of prevalence (World Health Organisation 1999; Mengarelli et al. 2006). Preprint: Dearlove and Wilson (2013).

subtype is also common in Italy, from where many people emigrated to central Argentina from the 1880s to 1920s (Mengarelli et al. 2006).

Out of the three epidemics with SIR dynamics, subtype 6a in Hong Kong had the highest support (PP=71.0%). The most recent ancestor was dated to between 1952 and 1962. The doubling time for this epidemic was much quicker than that for Argentina – estimated to be between 0.7 and 3.8 years. The epidemic was estimated to have peaked in 1986 with a credible interval spanning thirty years from 1963 to 1993. Subtype 6a accounts for 23.6% of all HCV infections in Hong Kong; this figure increases to 58.6% when considering infection in intravenous drug users (Zhou et al. 2006). The previous rarity of subtype 6a in China (Lu et al. 2005) has led to the suggestion that the HCV-6a epidemic was introduced from Vietnam, where it is dominant, through immigration (Zhou et al. 2011). The estimate of the most recent common ancestor from this new analysis is a little earlier than the immigration of the Vietnamese boat people in 1978-

Table 3.5.4: Epidemiological parameter estimates for SIR dynamics.

Parameter	2c Argentina	6a Hong Kong	6f Thailand
Intrinsic growth rate, r_0 (year^{-1})	0.139 (0.103, 0.193)	0.358 (0.182, 0.944)	0.271 (0.191, 0.507)
Doubling time, $\log(2)/r_0$ (years)	4.987 (3.591, 6.730)	1.936 (0.734, 3.809)	2.558 (1.367, 3.629)
Basic reproductive number, R_0	1.20 (1.04, 5.51)	1.44 (1.08, 13.17)	1.42 (1.07, 19.02)
Average duration of infectiousness, γ (years)	1.745 (0.274, 27.027)	1.244 (0.264, 14.085)	1.553 (0.275, 40.000)
Current prevalence, I_0 (%)	0.534 (0.021, 20.137)	0.191 (0.002, 6.235)	1.568 (0.051, 40.922)
Present effective population size, N_0 (10^3)	3.4 (1.4, 13.3)	0.2 (0.1, 0.8)	6.9 (0.9, 135.0)

1982 and 1987-1997 cited by Zhou et al. (2011); however it does concur with it being cited as a relatively recent epidemic by Tanaka et al. (2006).

There are many subtypes of HCV-6 found distributed around Asia. Subtype 6f seems to be restricted to Thailand, where again it is the most common causing 56% of HCV cases (Akkarathamrongsin et al. 2010). This analysis showed marginally greater support for the SIR model compared with the SIS model (PP=39.6% versus PP=38.1%). The difficulty in separating the two models is due to a very recent peak and subsequent deceleration of the epidemic. This is fitting with research showing a decline between 2005 and 2008 in intravenous drug use, a major risk factor for HCV infection in Thailand (Jatapai et al. 2010). The T_{MRCA} was dated to between 1955 and 1968, with the peak prevalence occurring between 1964 and 2008. The doubling time was found to be similar to that of subtype 6a in Hong Kong – between 1.4 and 3.6 years.

In all three datasets exhibiting SIR dynamics, an independent point estimate of prevalence was compared to the estimated prevalence from the genetic analysis, and shown by the intersection of the red lines in Figure 3.5.4. In all three cases, the independent estimate fell within the 95% credible interval of the trajectories. This suggests that in the case of SIR dynamics, genetic data alone can provide reasonable estimates of prevalence.

3.6 Discussion

This chapter uses a new metapopulation model of pathogen populations for combined epidemiological and genetic inference to understand the dynamics of 18 HCV datasets from across the globe, and to investigate the possible processes influencing genetic diversity. Somewhat surprisingly, there was no significant relationship between diversity and the intrinsic growth rate, r_0, after adjusting for the other factors (p-value: 0.6027). The time to the MRCA had the strongest predictive power out of the covariates included in the linear regression model (Table 3.5.2). This is explained by the rapid epidemic growth in all of the datasets, resulting in star-shaped genealogies due to branch lengths being long near the tips. In this scenario, the MRCA is only slightly younger than the epidemic, and it follows that diversity within an HCV dataset can be used as a good indicator of the age of an epidemic.

All except one of the epidemics (subtype 4a in Egypt) appear to have emerged in the last 100 years (Table 3.5.1). This result highlights the importance of 20th century phenomena such as blood transfusion and needle sharing in the spread of HCV (Pybus et al. 2003; Pybus et al. 2005). This analysis covers countries with a range of economic development, and the main transmission routes tend to differ according to wealth. In developed countries such as the United States and United Kingdom, intravenous drug use is the main risk factor; in economically developing countries transmission is more likely to have occurred through medical practices including surgery, haemodialysis, contaminated immunisation injections and other parenteral therapies (Pybus et al. 2003; Aslam et al. 2005; Shepard et al. 2005). Whilst the fact that these epidemics are recent is

not striking in itself (previously reported by Magiorkinis et al. (2009) and Shepard et al. (2005) amongst others), it is interesting that there is no obvious difference in dynamics according to route.

Type 1 was the most represented type in the meta-analysis (10 out of 18 datasets, Table 3.4.1), and all were found to have SI dynamics (Table 3.5.3). In only one case (Belgium subtype 1a) was there reasonable support towards the SIS or SIR models (SIS PP: 35.0%; SIR PP: 20.0%). The dynamical modelling used in this analysis shows that prevalence either reaches dynamic equilibrium (SIS), or falls as the number of susceptible hosts is exhausted (SIR). Why neither of these appear to have occurred in type 1 epidemics is difficult to pin down, though may depend on environmental, cultural and social factors relating to the viability of transmission. The differing dynamics of subtypes 1a and 3a in Belgium might hint at the effectiveness of treatment playing a part, as trials have shown that HCV types 2, 3 and 6 respond better to treatment than type 1 (Manns et al. 2001; Yu and Chuang 2009).

Three of the datasets showed SIR dynamics. Whilst they originate from different countries, and involve different subtypes, it is clear that these epidemics are linked in that they are globally rare, but locally dominant. In two of these datasets, there was also high support for the SIS model (PP=53.4% for the SIR versus PP=45.4% for the SIS in Argentina subtype 2c, and PP=39.6% versus PP=38.1% in Hong Kong subtype 6a). This suggests a recent peak in infection and subsequent slowing of the epidemic, though it is noted that whilst model selection singles out SIR dynamics as the best fit of the models

investigated, there may be other parametric forms that fit better, for example a step function. Despite the Subtype 6 epidemics in Hong Kong and Thailand having faster intrinsic growth rates than that in Argentina, similar estimates were obtained for all three datasets for the basic reproductive number (1.20-1.44), and mean duration of infectivity (1.24-1.55 years). Whilst there is considerable uncertainty in the estimates for both parameters, the basic reproductive number is noticeably lower than calculated previously. This can be attributed to the average period of infectivity being estimated to be much shorter than the 10-30 years used in the past (Pybus et al. 2001). The estimate of the duration of infectivity is surprising given that hepatitis C infection is chronic in 80% of people, and has lifelong infectivity (Lavanchy 2011). It is possible that the low estimates here allude to the majority of transmission occurring in the first year and a half of infection – perhaps representing the period before which a diagnosis is confirmed and any intervention in patient behaviour has taken place. In this scenario, we interpret the 'recovered' population not only as those who have cleared infection or have been removed from the model through death, but also as those who are no longer 'actively' infectious by interacting with those 'at risk' in the susceptible population (for example, by practicing safe and sterile needle use after diagnosis). This definition of 'recovered' individuals actually just being 'removed' and no longer partaking in the disease dynamics perhaps better represents the known biology of HCV. However, it should also be noted that the wide credible intervals are consistent with infectious periods of up to 27, 14 and 40 years for subtypes 2c, 6a and 6f respectively (Table 3.5.4).

Of course, there may be some bias in this estimation of R_0, as the detection of SIR dynamics is only possible once the epidemic has passed its peak. Moreover, this point is likely to occur sooner when R_0 is smaller. The results here suggest that the majority of epidemics have not yet reached that peak, and thus in time it is likely that the estimate will be refined as more epidemics begin to decelerate. However, there are shared features between the three datasets used to estimate R_0 here. The Hong Kong and Argentinean epidemics both appear to originate with migration events to the locale, and the Thai and Hong Kong epidemics have appeared recently.

For the analysis, a fixed mutation rate of 5.8×10^{-4} substitutions per year was used with the HKY85 substitution model (Hasegawa et al. 1985). This rate had been previously applied to a number of datasets included in the analysis. However, there is evidence to suggest that the true rate could be nearly double this, around 1.0×10^{-3} substitutions per site per year (Magiorkinis et al. 2009; Gray et al. 2011). Underestimating the substitution rate would cause the historical timeline of the epidemic to be systematically overestimated. This analysis also neglected to take into account possible heterogeneity in the molecular evolution rate between sites, codon positions, or branches of the tree. There has been work done suggesting that evolution in the NS5B gene follows a relaxed clock, and that evolution of the third codon position is relatively slower than the first two positions (Magiorkinis et al. 2009; Gray et al. 2011). Not taking this into account may mean that the estimates for some of the parameters and dates here are anti-conservative. Nevertheless, an advantage of implementing the analyses in BEAST is that this complexity can be incorporated in future work.

A model averaging approach was used to assess the fit of each of the endemic, SI, SIS and SIR models to each dataset. Whenever fitting models, it is important to be aware of balancing complexity with biological relevance. While it is clear that all three models represent simplifications of true epidemiological processes, they capture fundamental processes featuring in epidemics – including the exponential growth, equilibrium and deceleration phases. Although the SI model was preferred in 13 out of 18 datasets, there was little statistical support towards even the small elaborations of the SIS and SIR models (Table 3.5.3). It is possible that none of the five models included in the comparison adequately fitted the true population history of the data. In this case, it may make sense to include the Bayesian skyline plot (Drummond et al. 2005) alongside the parametric models being considered. This gives the option for the parametric models to be rejected, and may help inspire the development of novel, more realistic parametric models with the metapopulation approach.

Nevertheless, it should be noted that the Bayesian skyline plot only estimates the product of effective population size and generation time, and thus is difficult to directly interpret given that the effective population size is only directly proportional to the prevalence of disease at dynamic equilibrium (Volz et al. 2009; Frost and Volz 2010; Koelle and Rasmussen 2012). Whilst a skyline plot may have power to identify atypical and unanticipated trajectories in the effection population size over time, when appropriate, parametric models allow for epidemiological parameters to be estimated directly, making them more intuitive, easier to interpret and more statistically efficient. This is pertinent, because specialists in areas other than population genetics, such as

epidemiology and medicine, need to be able to easily interpret the results of these analyses.

The combined epidemiological and genetic approach outlined in Section 3.3 shows how it is possible to implement coalescent SI, SIS and SIR models for data analysis. However, there are a number of limitations to the method. Whilst it was found that the predicted prevalence from the SIR model was good when compared with independent estimates, the SIR model can only be fitted once the peak of the epidemic has been observed and thus is less useful when analysing an outbreak in real time. Calculating R_0 is also limited to the SIR model, as the transmission coefficient β_1 and rate of clearance of infection γ are confounded in the SIS model, and R_0 is undefined in the SI model. The intrinsic growth rate is a ready substitute for R_0, exhibiting equivalent threshold behaviour: when $r_0 \geq 0$ infection will persist in the population ($R_0 \geq 1$), and when $r_0 < 0$ the epidemic will eventually die out ($R_0 < 1$). The growth rate is well identified in the exponential growth phase of an epidemic, and can be used to compare two epidemics (for example to examine the effect of an intervention strategy); however it is not able to predict how the epidemic will unfold over time without other information such as the rate of recovery or present prevalence. As a result, the role of such modelling is to complement rather than replace traditional epidemiological approaches.

One of the main assumptions underlying the metapopulation coalescent is that the number of infected hosts is large. This assumption is consistent with the deterministic compartmental models used, but cannot be true at the start of an epidemic. One way of

overcoming this is to allow for stochasticity. Rasmussen et al. (2011) have used particle MCMC to fit a stochastic SIR model, though currently the genealogy is assumed fixed and ignores uncertainty around the true topology. An alternative to the coalescent might be to consider branching processes, though this method has so far been limited to simple birth-death processes (Leventhal et al. 2012; Stadler et al. 2012).

Theoretical work has shown that changes in contact networks affect epidemiological dynamics, and therefore genetic diversity (Campos and Gordo 2006; Volz 2008; Gordo et al. 2009; Koelle and Rasmussen 2012). This means that the assumption of random mixing is difficult to justify, despite it being commonly cited in compartmental modelling. It is possible that the variance in social mixing may be accounted for using a more general metapopulation formulation than used here (Wakeley and Aliacar 2001), which would allow for different classes of hosts. For example, this method could allow for the concept of a super shedder with a higher rate of infectivity to be modelled alongside standard host types. The assumption of negligible co-infection is harder to include, as co-infection means that recombination can occur. Evidence of recombination was seen in several HCV datasets from the literature review (Table 3.4.1), and these were simply excluded from this analysis. Whilst there have been attempts for recombination to be included in a population genetic framework (Bloomquist and Suchard 2010), it still remains computationally infeasible and such methods are rarely applied in practice.

The metapopulation basis to the model presented here is open to extension, and provides abundant grounding for future work. The host population size was fixed, but changes in demography could be incorporated. This might be desirable, for example, in the situation where data is available to analyse host and pathogen genetic diversity side by side. In this analysis, it was assumed that the samples in the datasets included contained no more than one sequence sampled per host. In this case it is easy to account for serial sampling using the standard method for tip dating implemented in BEAST (Rodrigo and Felsenstein 1999; Drummond et al. 2002). When there is more than one sequence per host, the collecting and scattering phases explained in Section 3.3.2 must be taken explicitly into account. These phases can approximately be taken to represent the within- and between-host evolution, and new apparatus would be required for inference.

Chapter 4: Evolution and Transmission of GII.4 Norovirus

4.1 Background to Norovirus

Noroviruses are positive-strand RNA viruses that form a genus in the *Caliciviridae* family – a family that also contains the genera sapovirus, vesivirus, and logovirus (Nilsson et al. 2003). Norovirus is most well-known for causing gastroenteritis in settings where individuals come into close contact – for example hospitals and nursing homes (Khanna et al. 2003; Calderon-Margalit et al. 2005; Schmid et al. 2005), schools and children's day care centres (Gallimore et al. 2004a; Button Gomez 2008), cruise ships (Widdowson et al. 2004; Vivancos et al. 2010), and military bases (Sharp et al. 1995; McCarthy et al. 2000). It affects people from all age groups, and has been detected worldwide throughout the year. The number of outbreaks tends to peak in the winter months - hence the alternative name 'winter vomiting disease'. However, in summer months of 2002 in England and Wales there was evidence suggesting a second peak in confirmed cases, disrupting the usual seasonal pattern (Lopman et al. 2003).

The first norovirus, Norwalk Virus, was identified in 1972, via immune electron microscopic (IEM) examination of experimentally infected volunteers after a recorded outbreak in Norwalk, Ohio in 1968 (Adler and Zickl 1969; Kapikian et al. 1972). The norovirus genome was first sequenced in the early 1990s, allowing a reverse-transcription polymerase chain-reaction (RT-PCR) method to be developed for diagnostic purposes (Jiang et al. 1990; Jiang et al. 1992; Ando et al. 1995). However, until recently, much work on the evolution and transmission of norovirus in humans has

been limited by the inability to grow in culture and lack of animal model resulting in a deficiency in virus stocks for experimentation (Atreya 2004). Whilst murine norovirus can be successfully grown in mouse dendritic cells and macrophages (Wobus et al. 2004), these cell lines have been shown to be unsuitable for growing human norovirus (Lay et al. 2010). Human duodenal tissues have been successfully infected with GII.4 strains, with foetal ileum cells showing limited viral replication after infection (Leung et al. 2010), but there is still a way to go until *in vitro* cell culture is possible (Papafragkou et al. 2013).

4.1.1 Structure and Evolution

Small, round and non-enveloped, noroviruses have a diameter of approximately 27-38nm, and a genome of length around 7.5kb (Jiang et al. 1990). The genome is split into three open reading frames (ORFs), as shown in Figure 4.1.1. ORF1 contains the code for non-structural proteins (including the RNA-dependant RNA polymerase, RdRp), whereas ORF2 and ORF3 code for VP1, a major capsid protein, and VP2, a minor capsid protein, respectively (Hardy 2005). The VP1 region shows the highest diversity within the genome, and is made up of the conserved shell region (S), and the protruding

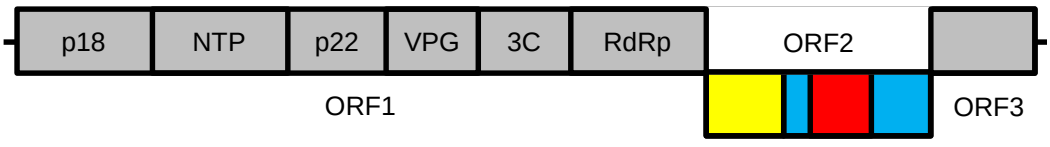


Figure 4.1.1: The norovirus genome. The norovirus genome is split into three open reading frames. ORF1 makes up the first two thirds of the genome, and is labelled with the non-structural replicase proteins. In ORF2, VP1 is made up of the shell (yellow), P1 (blue) and P2 (red). ORF3 codes for the minor structural proteinVP2. Adapted from Figure 1a in Donaldson et al. (2010).

domains P1 and P2. Of the protruding domains, P2 is found on the exterior of the capsid and contains a number of motifs that have a role in host cell binding and, as a result, viral antigenicity (Tan et al. 2004; Siebenga et al. 2007).

The norovirus genus shows high diversity, infecting a range of species, and is split into five main genogroups labelled GI to GV, as seen in Figure 4.1.2. Genogroups are further classified into subgroups using a point and number representing the subgroup, for example GII.4. Only three genogroups (GI, GII and GIV) have been found to infect humans (Koopmans 2008). However, these genogroups are not exclusive to humans - strains from GII are also found in swine (Wang et al. 2005; Oka et al. 2013), and strains from GIV in dogs and lions (Martella et al. 2007; Martella et al. 2008). Of the remaining two genogroups, GIII infects cattle and sheep (Oliver et al. 2003; Wolf et al. 2009), and

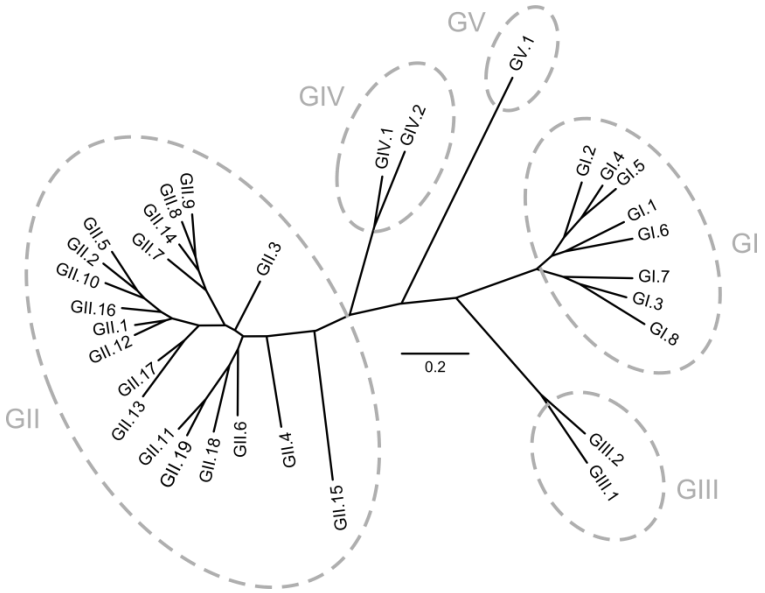


Figure 4.1.2: Diversity of norovirus. Phylogeny of 32 norovirus genotypes based on the VP1 sequence, showing the diversity of the main norovirus genogroups. Sequences from: Martella et al. (2008), Martella et al. (2007), Wang et al. (2005), Oliver et al. (2003), Karst et al. (2003), Fankhauser et al. (2002), Vinjé and Koopmans (2000), Sugieda et al. (1998) and Green et al. (1995).

GV is found in mice (Karst et al. 2003). Genotyping has generally been done in the past using two main sections of the genome: the polymerase gene in ORF1, and VP1 in ORF2, either through RT-PCR in the lab (Vinjé et al. 2004), or after sequencing using a BLAST search and tree-building (Kroneman et al. 2011). There is 45-60% divergence in amino acids across the genogroups, and 14-45% between genotypes (Zheng et al. 2006).

As in Hepatitis C, the high diversity found across norovirus genotypes is in part due to the fallible nature of their replication mechanism, the RNA-dependant RNA polymerase. Estimation of an evolutionary rate has so far been limited, with rates calculated between 1.9-8.98 $\times 10^{-3}$ substitutions per nucleotide per year using either just the capsid or the partial RdRp (Bok et al. 2009; Bull et al. 2010; Siebenga et al. 2010). Whilst not all mutations will result in viable offspring, some will enable new strains to emerge with phenotypes providing potentially better fitness, allowing these strains to co-circulate with or even dominate existing ones. Inter-genotypic recombination has also been reported to occur in the capsid gene and at the interface between ORF1 and ORF2 (Bull et al. 2005; Rohayem et al. 2005; Eden et al. 2013).

4.1.2 Pathogenesis and Transmission

Symptoms of norovirus include nausea, abdominal cramps and other muscle pain, vomiting and diarrhoea. The incubation period is relatively short, between 24 and 60 hours, and symptoms are generally cleared within 3 days (Lopman et al. 2003). The advised treatment is usually rest and drinking plenty of fluids to prevent dehydration; more severe cases may require electrolytes to be given intravenously (Hutson et al.

2004). Despite symptoms being resolved within the space of a few days, asymptomatic shedding of virions may last much longer (Aoki et al. 2010). Norovirus has a median infectious dose of 18 virions required for an adult to develop infection (Teunis et al. 2008). It is estimated that a single bolus of vomit contains 30 million virions (Caul 1994); so this coupled with prolonged shedding after infection and low infectious dose means that norovirus is highly contagious and easily spread through a number of routes.

Transmission most commonly occurs directly through contact with an infected individual via the faecal-oral route. Indirect transmission can be caused by contact with a contaminated surface or consumption of contaminated food, as seen by the food handler being the source of the outbreak described de Wit et al. (2007). Food and water may not necessarily be directly contaminated by an infected individual, for example filter feeders located near sewage and waste water outlets can carry the virus and be consumed (Nenonen et al. 2008; Schaeffer et al. 2013). If there is high prevalence of vomiting, aerosolisation may be attributed for transmission routes via infectious droplet contamination in the local environment (Jones et al. 2007; Repp and Keene 2012). It has also been shown that aerosolisation can occur with toilet flushing (Barker and Jones 2005; Johnson et al. 2013).

4.1.3 Susceptibility and Immunity

Histoblood group antigens (HBGAs) have been suggested to play a strong role in the susceptibility of humans to norovirus. HBGAs are carbohydrates found on the epithelial cells, and form binding and possible receptor sites for norovirus. There are three major

types of HBGAs that affect norovirus binding – ABO, secretor and Lewis – and it is suggested that they are recognised by different norovirus strains (Huang et al. 2003; Huang et al. 2005; Tan and Jiang 2005). The presence or expression of most HBGAs is due to the presence of the FUT2 gene in the human genome, which codes for fucosyltransferase. Around 20% of the population do not have the FUT2 gene, and this gives them immunity to GI.1 infection, though may allow infection with other genogroups (Lindesmith et al. 2003). In addition, it has been shown that individuals with blood group O are more likely to become infected with GI.1 norovirus (the subtype containing the originally identified Norwalk Virus), whereas those with blood type B are less at risk of becoming infected and showing symptoms (Hutson et al. 2002). GII.4 strains are thought to be highly prevalent (see Section 4.2.1) in part due to the fact they can bind A, B and O secretors, representing around 80% of the population (Debbink et al. 2012b). In contrast to GI.1, a study of GII.4 in China suggested that individuals with blood type A had an increased risk of infection, whereas blood type O had a decreased risk (Tan et al. 2008). Therefore understanding susceptibility of norovirus to humans remains complicated; whilst susceptibility to a specific norovirus strain may be restricted within the human population by, for example, blood type, the sheer diversity of strains and HBGA binding types mean that individuals are most likely susceptible to at least one circulating strain.

There is much debate on the length of immunity acquired after infection. It has been shown that GI.1 gives short-term immunity of at least a few weeks (Wyatt et al. 1974; Parrino et al. 1977). However, evidence towards long term immunity has been less

decisive, with subjects re-challenged with norovirus less conclusively immune after a six month period (Johnson et al. 1990; Atmar et al. 2011), and all individuals symptomatic within 27-42 months (Parrino et al. 1977). Bull et al. (2011) argue that there is some evidence for long term immunity for GII.3 norovirus, citing the slow evolutionary rate and limited number of variants in the last 40 years. They also argue that recent work supports the immunity lasting six to twelve months, and during this period there is sufficient time for a new antigenic variant to emerge. It is worth noting, however, that having immunity to one strain does not necessarily imply a lesser risk of being infected with a different strain of norovirus (Wyatt et al. 1974). All of the above have contributed to the difficulties in developing a norovirus vaccine (Donaldson et al. 2010).

4.1.4 Management in the Clinical Setting

Due to a general lack of timely diagnostic testing facilities, outbreaks of gastroenteritis in clinical environments due to norovirus are usually identified according to the Kaplan criteria (Kaplan et al. 1982a; Kaplan et al. 1982b). These are a set of clinical and epidemiological guidelines for diagnosing norovirus as the underlying agent as opposed to bacterial causes of gastroenteritis. The criteria suggest that the following are indicative of norovirus: vomiting in more than 50% of cases, an average of a 24-48 hour incubation period, an average duration of illness of 12-60 hours, and no bacterial pathogens isolated from stool culture. As genetic diagnosis (either by real time genomic sequencing or by RT-PCR) becomes mainstream in general practitioners' offices and hospitals, the reliance on this scheme will decrease. However, until that time comes,

Turcios et al. (2006) indicate that the Kaplan criteria are the strongest method of identifying norovirus outbreaks.

The cost of norovirus outbreaks to health services can be huge, and are estimated to cost the National Health Service (NHS) in England £115 million each year (Lopman et al. 2005). A single outbreak in 2002 in Eastbourne was estimated to have cost £279,115, with lost bed days, cancelled surgery, staff sickness and cleaning of the local environment being among the most expensive factors (Cooke et al. 2003). As with any nosocomial pathogen, cleaning of hospital wards is important in reducing transmission of norovirus. Noroviruses are environmentally stable; they can survive in temperatures between 0 and 60°C, concentrations of chlorine up to 5000ppm, ethanol solutions less than 70% and a wide pH range (Barker et al. 2004; Duizer et al. 2004). This is likely due to the lack of lipid envelope in norovirus, therefore making it less affected by lipophilic chemicals at the concentrations used for disinfection (Springthorpe and Sattar 1990). In addition, alcohol hand sanitizers, used both in hospitals and in the community for rapid, accessible disinfection, have been shown to be less effective against norovirus than other pathogens, with hand washing a better alternative where viable (Liu et al. 2010). As a result, clearing virions from an affected area quickly is difficult, especially in a busy clinical setting where staff, patients, visitors and equipment move between wards. To aid with the control of norovirus in hospitals, current practice involves patient isolation, ward closure and restricting visitor access (Greig and Lee 2009; Norovirus Working Party 2012). The impact of such intervention strategies and deep cleaning is unknown

(Weber et al. 2010), so understanding the main transmission routes and sources of infection of norovirus is of great importance.

4.2 Motivation

Of the 19 norovirus GII subgroups, GII.4 has been the most dominant since the early 1990s, despite continued co-circulation of other genotypes. In Europe, the Foodborne Viruses in Europe Network estimate that GII.4 represents over 83% of reported norovirus cases in hospitals and residential institutions between 2001 and 2006, and almost 50% of all other outbreaks including those occurring in schools and from catering (Kroneman et al. 2008). Further, the genotype has caused up to 80% of outbreaks and sporadic cases globally and is the cause of at least five major documented pandemics in recent years (Bull and White 2011). This means that GII.4 is a key target for vaccine and therapy development (Parra et al. 2012). Insights into norovirus microbiology such as the replication process, protein structure and function, and interaction with the host immune system have been problematic due to the inability to culture (Atreya 2004). However, until these methods catch up, detailed investigation of the genetic epidemiology of disease can begin to untangle both the evolution and transmission of disease side by side (Grenfell et al. 2004), to improve the understanding of norovirus microbiology from a second perspective and predict the effects of the microbiology of norovirus.

4.2.1 Current Understanding of GII.4 Norovirus

For ease of identification, new and emerging norovirus strain types are named after the place and year that they were first isolated or confirmed. The first well-documented GII.4 strain emerged in the USA in 1995-6 and persisted into the late 1990s (Noel et al. 1999). In 2002 a new strain, Farmington Hills, spread throughout Europe, causing an increase in year round outbreaks, in contrast to the general winter peak of prevalence (Widdowson et al. 2004). Further global outbreak strains have included Hunter 2004 (Bull et al. 2006a), Den Haag 2006b (Tu et al. 2008; Eden et al. 2010), and New Orleans in 2009 to 2010 (Yen et al. 2011). There are also many other localised strains, which are notably divergent from the common global strains, but appear to be more geographically limited.

There have been a number of papers inspecting the rapid evolution of the GII.4 major capsid proteins (ORF2) between the major epidemic strains mentioned above (Siebenga et al. 2007; Lindesmith et al. 2008; Bok et al. 2009; Bull et al. 2010; Siebenga et al. 2010). These have shown that the majority of mutations the have occurred in the P2 subdomain, which is located on the exterior of the capsid, altering the antigenicity. Lindesmith et al. (2011) further showed that there has been a significant amount of evolutionary change in the antibody epitopes between major strains (for example, showing that antibodies from the 2006 GII.4 pandemic strain failed to recognise GII.4 virus-like particles from 2007 and 2008). They conclude that this was most likely to be in response to herd immunity – a view reiterated by Debbink et al. (2012a). However, these studies only focus on a small fraction of the genome, with ORF2 representing only

approximately 1,600b, or 21%, of the whole norovirus genome (Hardy 2005). Large-scale full genome studies have so far been limited, with very few whole GII.4 genome sequences available until the last three to five years (Siebenga et al. 2010).

Since the spring of 2012, a new strain has spread worldwide – Sydney 2012. This strain is believed to have caused one of the biggest winter outbreaks of norovirus in the UK, and has been isolated in locations across the globe including Austria, China, Denmark, Italy, Switzerland, and USA (Fonager et al. 2013; Giammanco et al. 2013; Huttner et al. 2013; Mai et al. 2013; Maritschnik et al. 2013; van Beek et al. 2013). In the UK, the Health Protection Agency (now Public Health England) reported a 49% increase in infection compared to the average number of cases over the previous five years, with a 44% increase in the number of hospital outbreaks recorded (Adams 2013). The majority of this increase was due to an earlier than usual peak in cases – with 356 outbreaks reported between October and November 2012, compared to 140 reported during the same period in 2011 – and this dominance over the previously circulating strain is in line with estimates of increased prevalence from the emergence of other GII.4 pandemic strains (Hall et al. 2013). Over the Christmas period, hospital outbreaks garnered much media attention, since 194 wards were closed during the last four weeks of 2012 due to control measures, restricting visitor access (Health Protection Agency 2013).

4.2.2 Hospital Transmission

Ward closures and restriction of visitors are common practice during clinical outbreaks of norovirus (Greig and Lee 2009; Norovirus Working Party 2012), with the main aim of

reducing transmission to other parts of the hospital. Infection can take place in a number of ways; the patient may have entered the hospital already infected, or was infected in the hospital through direct person-to-person transmission (patient, staff or visitor) or environmental contamination. Identifying which of these scenarios has occurred is problematic, yet this information could inform improved methods of intervention when outbreaks occur. For example, closing wards and restricting movement around the hospital (leading to the cancellation of, for example, scheduled surgery (Cooke et al. 2003)) might prove unnecessary if a large outbreak was in fact due to a number of isolated independent introductions in to the hospital with very little onwards infection, and not due to rapid transmission of a single strain throughout individual wards.

Partial sequencing of select regions of the norovirus genome have been used previously in outbreak analysis to link cases (Dingle and Norovirus Infection Control in Oxfordshire Communities Hospitals 2004); however the resolution of whole genome sequencing allows the possibility of tracking transmission at a finer scale. Recent work tracing the spread of an outbreak via the accumulation of mutations across the genome shown that most *Clostridium difficile* infection cannot, in fact, be explained by direct transmission from symptomatic patients within the hospital (Eyre et al. 2013; Walker et al. 2012; Cule and Donnelly, submitted), and has been used to show the directionality of transmission in five immunodeficient patients with norovirus (Kundu et al. 2013). In this chapter, I go beyond previous studies of norovirus transmission by investigating norovirus outbreaks using whole genome sequencing in Oxfordshire hospitals over a

four year period, in order to address the question of the main sources of norovirus infection in the hospital setting.

The Oxford University Hospitals (OUH) NHS Trust is made up of four hospital sites, and annually covers over 670,000 outpatient appointments (Oxford University Hospitals NHS Trust 2012). In 2009, my colleagues in the Modernising Medical Microbiology Consortium initiated a study into the transmission and evolution of norovirus in OUH. One of the key features of this study is the concurrent collection of both genetic and epidemiological data, the latter accessible via infrastructure put in place by the Infections in Oxfordshire Research Database (IORD). This database contains anonymised hospital records, including admission and ward movements, alongside laboratory information for any collected samples (Finney et al. 2011). This represents an extraordinary resource for a single NHS trust, and a unique opportunity for investigating norovirus transmission dynamics via a combined genomic and epidemiological approach.

4.2.3 Chapter Aims

In this chapter, I investigate GII.4 norovirus at three levels. I first investigate GII.4 norovirus globally, to calibrate the rate of evolution using whole genome sequencing. This is fundamental to detecting transmission through a comparison of norovirus genomes, because the approach relies on the interpretation of genetic relatedness in terms of recent transmission, something that is assisted by an estimate of the real-time rate of genomic divergence. Secondly, I look at genetic diversity within a single season,

to test whether the unusual dynamics of the winter 2012-2013 season are reflected in norovirus genomes isolated during the same period. These both lead to the main focus of this chapter, where I use a stochastic model to combine genetic data with electronic patient records to test the hypothesis that the majority of patient transmission in Oxfordshire hospitals over four seasons from 2009 to 2013 is due to infection acquired from another patient within the hospital.

4.3 Methods

4.3.1 Evolutionary Analyses

To investigate the evolution of GII.4 norovirus, I downloaded all norovirus sequences that were at least 7kb in length and had a known year of sampling from GenBank. I then typed these using the NoroNet genotyping tool (Kroneman et al. 2011), and I discarded any sequences found to be from subgroups other than GII.4. This resulted in 401 sequences (Appendix C), which I then aligned using MUSCLE under the default parameters (Edgar 2004). I also obtained a smaller dataset of only a single reference genome per major strain type of GII.4 as defined in Eden et al. (2013) and aligned in the same way (these sequences are indicated with an asterisk in Appendix C). I inferred norovirus phylogenies using PhyML 3.0 (Guindon et al. 2010), under the HKY85 model of nucleotide substitution (Hasegawa et al. 1985).

4.3.2 Winter 2012-2013 Dataset

Between October 2012 and January 2013, colleagues collected samples from across the United Kingdom and Jersey (Figure 4.3.1). This specific period covers the unusually

early arrival of norovirus in the UK, before the new Sydney 2012 strain responsible had been identified and reported in the literature (van Beek et al. 2013). Samples were taken from patients with symptoms of vomiting or diarrhoea in hospital, and community outbreaks identified as having likely been caused by norovirus using the Kaplan criteria (Kaplan et al. 1982b).

RNA from the whole stool samples were sequenced, since norovirus is non-culturable and therefore difficult to isolate, using the protocol in Batty et al. (2013). In that paper, mapping was used for assembling the genomes after sequencing to distinguish norovirus sequence reads from those of other organisms within the sample. However, the extensive genetic diversity between norovirus strains makes mapping-based

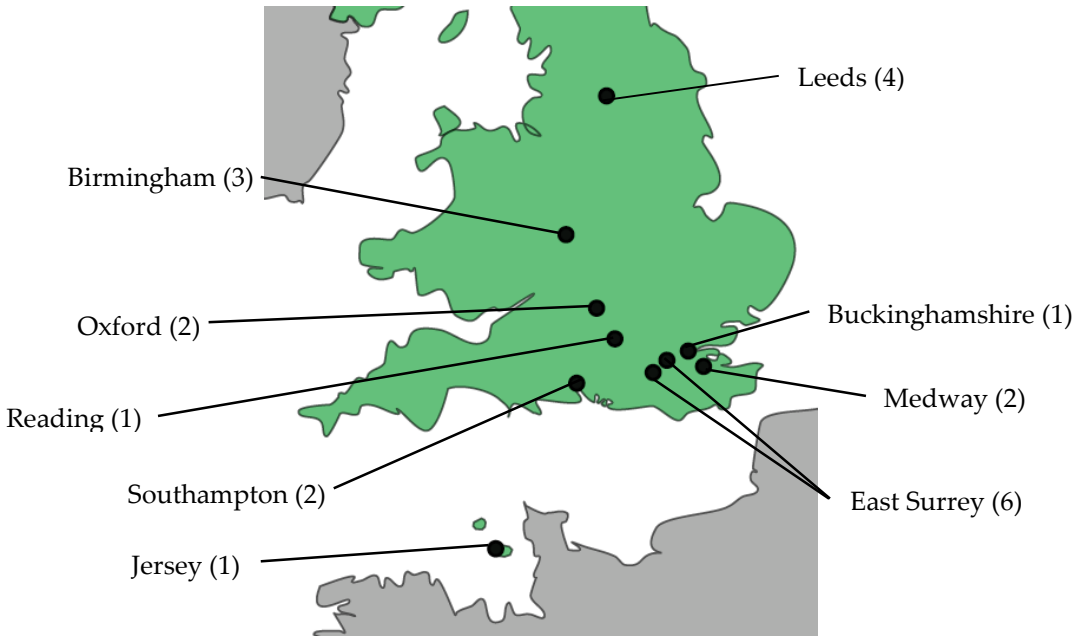


Figure 4.3.1: Norovirus sampling locations for the winter 2012-13 dataset. Map showing the locations from which samples were obtained during the winter 2012-13 outbreak. The value in brackets next to each location indicates the number of samples from both the community and hospital that could be adequately sequenced and used in further analysis.

approaches problematic due to their dependence on a specific reference genome chosen to map against – for example, it was found that sequencing reads from a New Orleans 2009 type strain did not map well to the Sydney 2012 reference strain (accession number JX459908). With this dataset, there could have been new, distinct variants and therefore a more robust *de novo* assembly pipeline was developed by Dr David Wyllie to remove the need for mapping to multiple references and for better identification of recombinant sequences (Wong et al. 2013). Firstly in his approach, for each sample, all sequenced reads were compared to 91 norovirus reference sequences using blast (Wong et al. 2013). Only reads with a good match to at least one of these reference sequences were kept, and these were then *de novo* assembled using VICUNA (Yang et al. 2012). This resulted in 22 full genome sequences covered by a single contig, which I aligned using MUSCLE (Edgar 2004) under the default parameters prior to further analysis. The reads for these 22 sequenced are deposited in the European Nucleotide Archive Sequence Read Archive under study accession number PRJEB4318 (minus C00014392).

4.3.3 Oxford Norovirus Outbreak and Patient Data 2009-2013

Outbreaks of norovirus in OUH are identified according to the Kaplan criteria (Kaplan et al. 1982a; Kaplan et al. 1982b). The criteria include vomiting in more than 50% of cases, an average incubation period of 24 to 48 hours, an average duration of illness of 12 to 60 hours, and no bacterial pathogens isolated from stool culture (implying that the outbreak is likely to be caused by norovirus). Standard protocol involves diagnosing samples using an immunoassay test, which gives a positive result in less than twenty minutes, although only up to a maximum of six samples per identified possible

outbreak are tested until the first positive result. Since these tests have poor sensitivity, samples are also sent for PCR at the reference laboratory.

Outside of this standard protocol for identifying patients with norovirus, stool samples were obtained for whole genome analysis whenever possible in collaboration with the nursing stuff. Samples were whole genome sequenced and *de novo* assembled using the methods described for the winter 2012-13 isolates in Section 4.3.2. For inclusion in the analysis, I required that sequences were at least 7.5kb in length and assembled into a single contiguous sequence. In the case of more than one isolate being whole genome sequenced for a patient, the sample with the earliest date was used for further analysis; if multiple samples were taken on this date, then one was chosen at random. I aligned sequences using MUSCLE (Edgar 2004). Since the ends of the contigs tended to be of low quality, I trimmed the alignment to include only the three open reading frames.

From September 2009 to July 2013, any patients that had norovirus infection confirmed via immunoassay, PCR or whole genome sequencing methods were automatically recorded in the IORD database as being symptomatic. Any unconfirmed patients, but with symptoms of diarrhoea and vomiting most likely due to norovirus – for example, due to known contact with someone who has confirmed infection, or is known not to have bacterial cause of infection – were also recorded in the extract from IORD.

For each symptomatic patient, epidemiological data about the hospital admission during which they had norovirus were obtained from IORD by Prof Tim Peto, curator of

the database. These data included the day on which symptoms first appeared, the date of any confirmed test results and genetic samples, the specific wards that patients visited, and the start and end times of these ward stays. Any patients for which ward stay information was unknown were removed from the analysis, as were those with obvious database errors such as admission times chronologically later than discharge times. These types of errors are inevitable in large databases such as those for a whole NHS trust, and are therefore just deemed as missing at random.

For the stochastic transmission model (see next section), information is required about the number of susceptible patients in the hospital at any point in time, in addition to those who are infected. Only wards visited by a symptomatic patient are relevant to the analysis, as susceptible patients in other wards would not have had the potential for transmission in the same way as those who shared wards with an infectious individual. For these wards, the only data obtained from IORD were the admission and discharge times for each patient that had visited.

4.3.4 Bayesian Analysis of Transmission in OUH

To investigate patterns of transmission in OUH, I used a stochastic transmission model developed by Dr Madeleine Cule and Prof Peter Donnelly (Cule and Donnelly, submitted), adapted for norovirus. In what follows, I give a brief overview of the original model which was used to investigate the epidemiology of *C. difficile* transmission, along with details of how this was adapted to take into account the known and distinctive biology of norovirus.

Original Model for *C. difficile*

Cule and Donnelly's model is based on a stochastic SIR model; such models have previously been applied to small datasets such as one or two wards, as in Forrester et al. (2007) and Cooper et al. (2008). The different states of infection that a patient is assumed to progress through are shown in Figure 4.3.2a. Initially, all patients are assumed to be susceptible (S). After infection, a susceptible patient first becomes colonised (C), a state which represents the incubation period before a patient becomes symptomatic and therefore infectious. The rate at which this occurs depends on whether the patient has already been admitted to hospital. The infectious period is split into three states: I_1 is the infectious population that have not yet had a confirmed diagnosis, I_2 corresponds to those infectious within a fixed window around the diagnosis time, and I_3 the infectious population after the test has taken place, aiming to capture the effect of patient isolation.

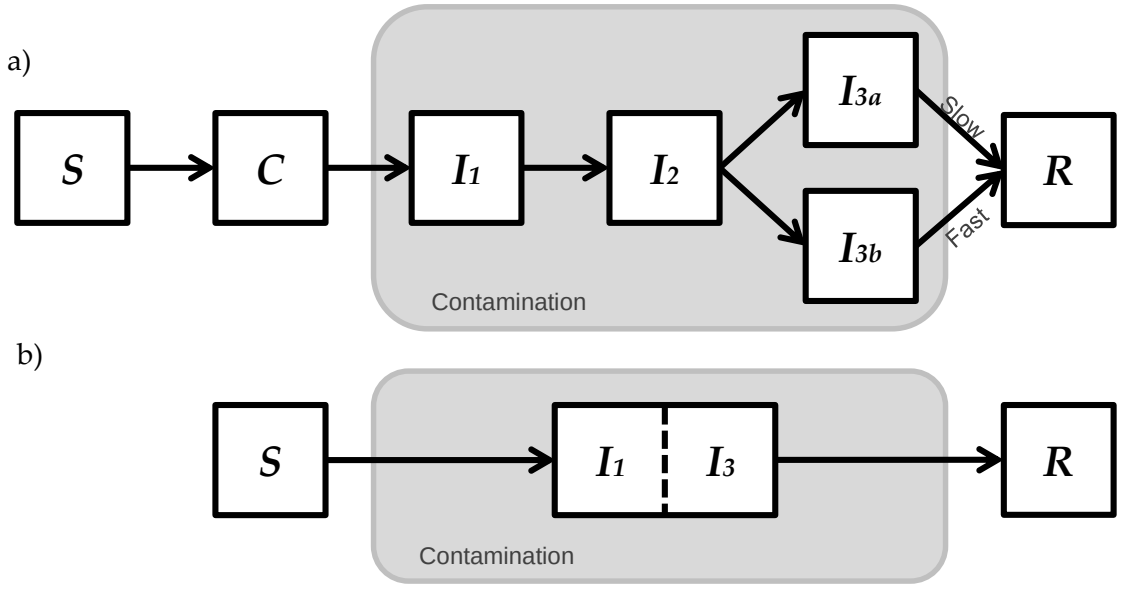


Figure 4.3.2: Stochastic compartmental SIR model. Patients are classed as either susceptible (S), colonised (C), infected (I) or recovered (R). In the original model for *C. difficile*, the infected compartment is split into pre-test (I_1), near-test (I_2) and post-test (I_3), and recovery can either be slow or rapid. This is simplified in the norovirus model, retaining only a single infectious state with a record time to test. Adapted from: Cule and Donnelly (submitted).

Random mixing is assumed among individuals in the same ward, so that any patient is equally likely to infect any other. Outside of wards, patients are considered equally likely to infect patients in the same hospital. Recovery (and subsequent removal from the model) can either be fast or slow, depending on the length of infectiousness (labelled as I_{3a} and I_{3b} respectively in Figure 4.3.2a). Definitions for all parameters associated with moving through these states are given in Table 4.3.1.

Cule and Donnelly developed a Gibb-sampling MCMC approach to estimate the model parameters in which the source of colonisation (infection in our simpler model) is tracked for each patient. This can either be from the background, either inside or outside the hospital, or due to another patient in the hospital. Transmission can only be inferred as compatible between two patients if the genetic sequence types (ST) of samples isolated from the two patients are the same. When this is true, the source is then broken down into the ward stay during which the transmission event took place, and the transmission route. In the full model, there are therefore nine possible patient-patient routes: contact on the same ward (either directly, or environmental), contamination attributed to a patient after they have left a ward, and within the hospital (but not in the same ward) for each of the pre-test (I_1), near-test (I_2) and post-test (I_3) infectious states.

Stochastic Model for Norovirus Transmission

The model in this chapter was adapted by Dr Cule for norovirus, as depicted in Figure 4.3.2b. The infectious states around the test are much simplified, allowing for only a single rate of infectiousness rather than rates varying according to whether the patient is

Table 4.3.1: Summary of parameters in the stochastic SIR model.

Parameter	Description
Rates of transmission	
* β_0	Background rate of infection in hospital
* β_1	Rate of patient-to-patient transmission within a ward (no difference between pre- or post-test)
β_2	Rate of transmission due to post-discharge contamination
* β_3	Rate of colonisation when outside hospital
β_4	Rate of transmission due to direct contact before $T-a$, where T is the test time and a is fixed
β_5	Rate of transmission in an interval $[T-a, T+b]$, where T is the test time and a, b are fixed
β_6	Rate of transmission after time $T+b$ but before recovery, where T is the test time and b is fixed
Multipliers	
α_i	Relative infectiousness of sequence type i
η	Relative rate of transmissions between wards compared to within wards
* ψ	Relative infectiousness of contamination compared to direct contact
Rates of disease progression within a patient	
* λ_1	Time from colonisation to test, with no incubation period
λ_{1a}	Time from colonisation to start of infectious period, inside the hospital
λ_{1b}	Time from colonisation to start of infectious period, outside the hospital
λ_2	Time from colonisation to start of infectiousness
λ_{2a}	Time from colonisation to onset of infectiousness, inside the hospital
λ_{2b}	Time from colonisation to onset of infectiousness, outside the hospital
λ_3	Time spent in I_1
* μ_1	Duration of recovery (single rate for all patients)
μ_{1a}	Duration of recovery (for slow recovering patients)
μ_{1b}	Duration of recovery (for fast recovering patients)
* μ_2	Duration of post-discharge contamination
Other	
θ	Proportion of fast recovering patients
* ϕ	Probability of being colonised on admission

* Only these parameters were included in the norovirus analysis.

pre-, near- or post-test. This is sensible, since there is little to suggest the rates of infection attributed to a particular patient differ according to whether or not norovirus has been confirmed – unlike *C. difficile*, norovirus treatment does not require specific drugs targeting the pathogen, instead focusing on rehydration. Further, testing is not consistently performed across patients and so a true test date is unknown in most cases. However, for presenting the results, I used the date of first confirmed symptoms as the nominal test date, and, where this information is not recorded but a sample was taken for analysis, the sample collection date. Consistent with this change, the time from colonisation to infection, λ_1, is now interpreted as the time it takes for symptoms to be recorded after initial infection, and the total duration of infection is given by the reciprocal of the total infection rate due to all possible routes of infection. Finally, a single rate of recovery is assumed. This simplified model resembles the deterministic SIR model applied in Chapter 3.

In the stochastic model, genetic information is incorporated alongside the epidemiological model in the form of a sequence type (ST); transmission can only be inferred between two patients if they are infected by the same ST. In order to apply the stochastic model to norovirus, it was necessary to define norovirus STs, a concept that is not routinely used in the norovirus literature. All strains were the same genogroups (GII.4), the lowest taxonomic level in common usage in norovirus. Therefore I defined groups of similar sequences using a clustering algorithm, whereby each isolate is a maximum number of single nucleotide polymorphisms (SNPs) from at least one other member of the same cluster. This is equivalent to finding the unrooted phylogeny of all

sequences, and then cutting all branches which have a length greater than the SNP cut-off value, and defining clusters according to the subtrees of isolates remaining. Choosing an appropriate cut-off is difficult, and it might be reasonably expected to have an effect on the results, so I implemented models using thresholds of 1, 3 and 10 SNPs, which we will see correspond to 9, 28 and 91 days of norovirus evolution respectively, to investigate the robustness of results.

The stochastic SIR model assumes that the rates of colonisation, infection and recovery are constant. Clearly this may not be the case for norovirus infection over a period of several years, due to the known periodicity of a peak in winter cases and very little infection taking place over the summer (Mounts et al. 2000). Therefore, I split the four years of data into seasons to allow the rates of transmission to be estimated independently and compared across years. I defined a season as starting at midnight on 1st September and ending at 11:59pm on 31st August the following year, as this point falls during the period of fewest reported cases of norovirus in the UK (Health Protection Agency 2013).

I ran two independent MCMC chains for each combination of SNP threshold and season, starting from different initial values to check for convergence. I ran the MCMC for a chain of 500,000 steps, with samples taken every 500 iterations and the first 10% of samples discarded as burn in. All of the priors are assumed to be gamma distributed, to facilitate the Gibbs sampling algorithm, and are given in Table 4.3.2. The gamma distributions assumed *a priori* for λ_1, $\mu_1$and μ_2 all have a median of five days, as this is a

conservative estimate of duration of symptoms to occur, recovery period and contamination respectively, given what is currently known about the epidemiology of norovirus infection (Weber et al. 2010). Results quoted use the median for point estimates, and the 2.5% and 97.5% quantiles for credible intervals.

Table 4.3.2: Scale and shape parameters for the gamma distributed priors.

	Gamma prior parameters	
Parameter	**Scale**	**Shape**
$\beta_0, \beta_1, \beta_3$	1.00	0.001
ψ	1.00	1.00
λ_1	2.31	0.10
μ_1, μ_2	1.38	0.10

4.4 Results

4.4.1 Evolutionary Relationship between Global GII.4 Strains

The phylogeny of the dominant GII.4 strains since the 1970s reveals an interesting dynamic of recurrent strain emergence (Figure 4.4.1a). Unlike, for example, the ladder-like trees associated with seasonal influenza virus (Bedford et al. 2011), new norovirus strain types do not appear to emerge from the most recently circulating variation. Instead, new strains seem to be more closely related to the major strains from at least two seasons previously. For instance, the most recent Sydney 2012 strain (and related Hong Kong 2012 strain, isolated around a similar time (Chan and Chan 2013)) is far more closely related to the Osaka 2007 and Asia 2003 strains, than the previously circulating New Orleans 2009 strain. Another such example includes Den Haag 2006b being more closely related to strains that originally emerged in 2001-2, than to Hunter

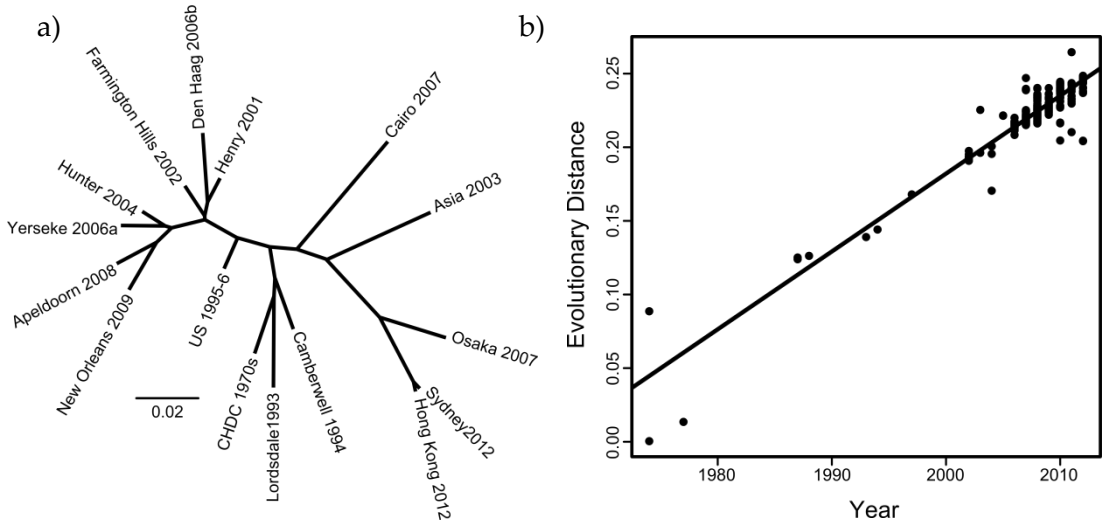


Figure 4.4.1: Evolution of GII.4 Norovirus. a) Maximum likelihood tree showing the ancestral relationships of 16 major GII.4 strains. b) Scatterplot showing the evolutionary distance in expected substitutions per site between 400 GII.4 norovirus sequences and an ancestral CHDC strain from 1974. The gradient of the regression line represents the evolutionary rate in substitutions per site per year.

2004, Yerseke 2006a or either of the strains from 2007. This suggests that old strains succeed in persisting despite a dominant new strain, evolving phenotypes that give them a distinct advantage over the currently circulating strain, results that update and extend the observations of Bull and White (2011) on the VP1 capsid gene.

Despite the punctuated emergence of new strains over time, the rate of molecular evolution over the last 40 years has been remarkably clock-like. In Figure 4.4.1b, the expected evolutionary distance of 400 GII.4 sequences (Appendix C) to an ancestral CHDC strain from 1974 (GenBank accession number: FJ537134; Bok et al. (2009)) is plotted against time. The gradient of the regression line estimates the mutation rate in GII.4 norovirus to be 5.29×10^{-3} substitutions per site per year, with a nominal 95% confidence interval of $(5.09, 5.48) \times 10^{-3}$. Note that this confidence interval is only a

guide to the uncertainty in the estimate of the mutation rate, due to the non-independence caused by shared evolutionary history.

Assuming a genome length of 7,560b, the genome-wide mutation rate is approximately 3 mutations per month, or 1 mutation every 9 days. It is worth noting that the majority of isolates that I downloaded from GenBank were sampled from 2006-2008. To investigate the influence of this apparent oversampling, I resampled sequences from these years at a lower frequency and re-calculated the gradient of the regression line. Reassuringly, no significant change in the estimate was seen (p-value ≤ 0.0001).

4.4.2 Circulating Strains in England and Jersey, Winter 2012-2013

Prior to the emergence of the Sydney 2012 strain, the previously dominant global GII.4 lineage was the New Orleans 2009 strain (Yen et al. 2011; van Beek et al. 2013). These two lineages are relatively divergent, differing at 803 nucleotides, or 11% of the genome. To gauge the relationships between recent local norovirus outbreaks and the globally circulating strains, I compared 22 newly-sequenced norovirus genomes from England and Jersey isolated during winter 2012-13 to New Orleans 2009 and Sydney 2012, focusing on the sites that distinguish the two reference genomes. For each sequence on the y-axis, Figure 4.4.2 shows whether each site along the x-axis is the same as the Sydney 2012 reference in blue, the New Orleans 2009 reference in yellow, or different to both references in turquoise. Two of the 22 genomes, Jersey 10/Dec/2012 and Southampton 21/Dec/2012, are clearly more closely related to the New Orleans 2009 reference strain, and 18 more closely resemble the Sydney 2012 reference strain.

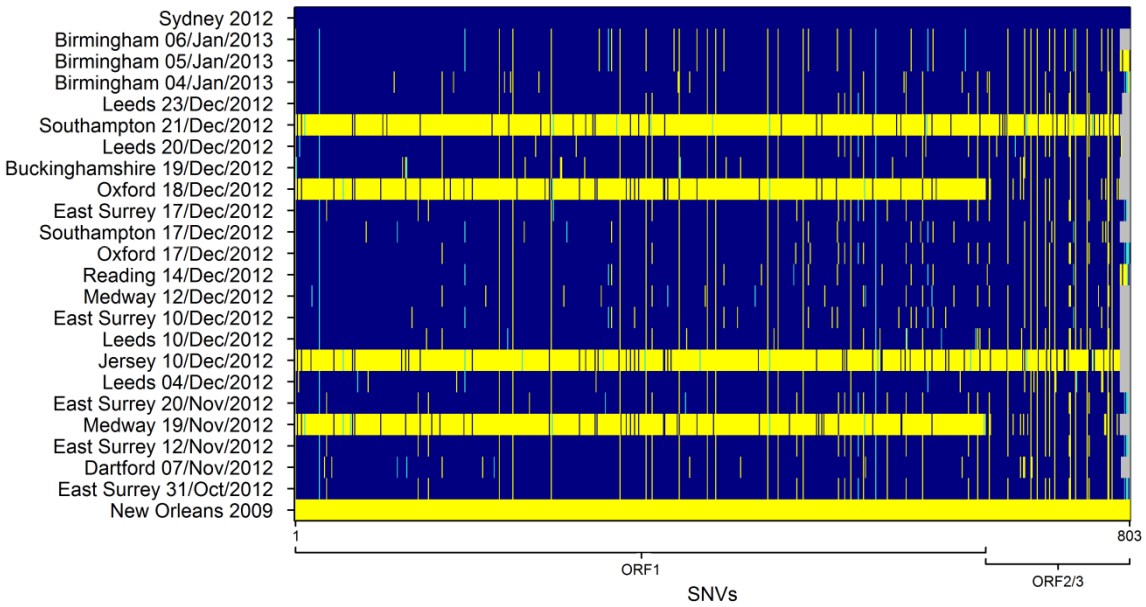


Figure 4.4.2: Comparison of isolates from winter 2012-13 with the last two most recent global strains. Graph depicting all 803 single nucleotide variants between New Orleans 2009 and Sydney 2012 along the x-axis, and the UK Winter 2012-13 samples and reference strains along the y-axis. The relative positions of the ORFs have been shown for reference. Nucleotides identical to the Sydney 2012 variant are displayed in blue, whilst nucleotides identical to New Orleans 2009 are depicted in yellow. Turquoise indicates variants that that differ from both Sydney 2012 and New Orleans 2009. Grey depicts bases that were not called.

Interestingly, two sequences (Medway 19/Nov/2012 and Oxford 18/Dec/2012) appeared to be New Orleans-Sydney hybrids, as judged by their mosaic genomes in which the majority of sites match those in New Orleans 2009, but where sites towards the end of the genome match the Sydney 2012 reference strain. This mosaicism implies that the two sequences represent recombinant sequences with a recombination breakpoint around the ORF1/ORF2 overlap, the same location found by Fonager et al. (2013).

Evidence of phylogenetic incongruence between ORF1 and ORF2/ORF3 (Figure 4.4.3) supports this observation. The New Orleans 2009-like and Sydney 2012-like strains are found in monophyletic clusters (indicated by yellow and blue, respectively) in both the ORF1 and ORF2/ORF3 trees. The two previously identified recombinant sequences,

Medway 19/Nov/2012 and Oxford 18/Dec/2012, are found together in both trees, as shown in red. However, in ORF1 these sequences cluster with New Orleans 2009, and in ORF2/ORF3 they are most closely related to Sydney 2012, confirming a difference in ancestry between the open reading frames.

Notably, there is no obvious spatio-temporal structuring in the phylogenies in Figure 4.4.3. Sequences similar to both the Sydney 2012 and New Orleans 2009 strains are isolated throughout the October 2012-January 2013 period, and the Sydney 2012 strain is widely spread. In Southampton, both Sydney 2012-like and New Orleans 2009-like sequences were found within a four day period from 17th December 2012 to 21st

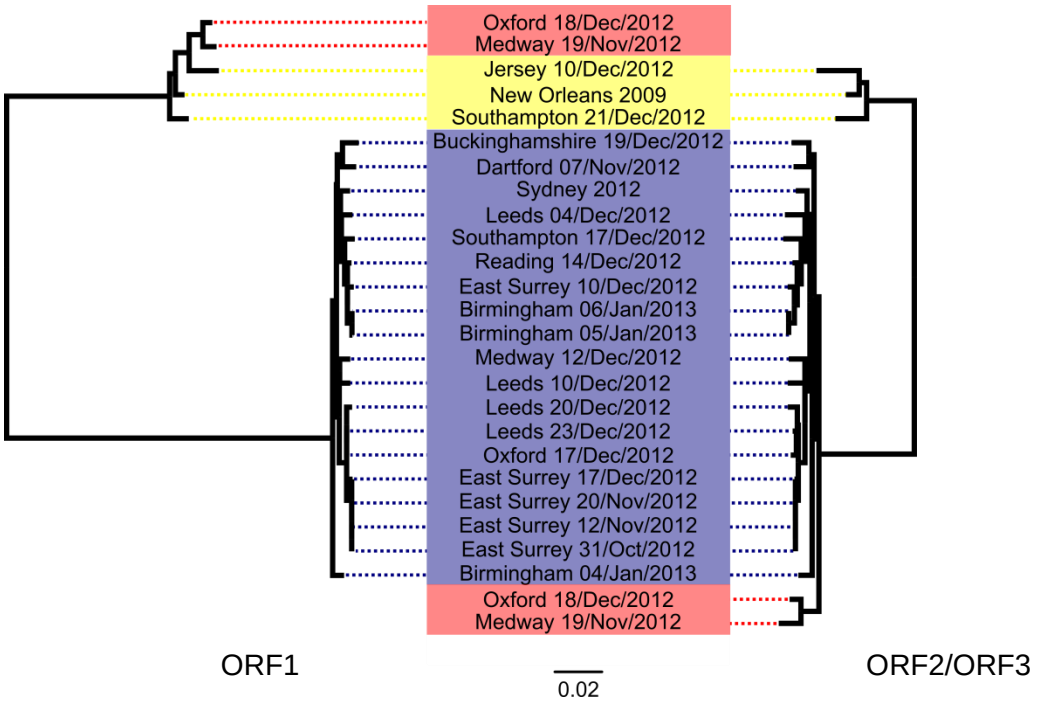


Figure 4.4.3: Ancestral relationships in ORF1 and ORF2/3 of winter 2012-13 isolates. Mid-point rooted maximum likelihood trees for ORF1 (left) and ORF2/ORF3 (right) regions of the genome for the 22 isolated and the two reference sequences. Clades are coloured by the reference sequence with which they cluster: New Orleans 2009 is shown in yellow and Sydney 2012 in blue, as in the previous figure. The two sequences indicated in red switch clusters according to ORF, and thus are deemed recombinant.

December 2012. The two recombinant sequences were sampled a month and around 280km apart, and despite having evidence for a common recombination breakpoint, have 85 substitutions between their genomes. Further, there is remarkable diversity in the Sydney 2012 strain type within a single location. In Birmingham, three sequences were sampled over a three day period. Between those sampled on the 5th and 6th January, there is only a single nucleotide difference; however, there are at least 98 differences between those and the isolate from 4th January. Such divergence is unlikely through mutation alone using the rate calculated in Section 4.4.1, so this suggests that there were multiple introductions of the virus in Birmingham.

4.4.3 Epidemiology of Norovirus in Oxfordshire Hospitals

The number of patients infected with norovirus varied widely in Oxfordshire hospitals across the four seasons covered, from 2009 to 2013 (Table 4.4.1). The highest number of cases attributed to norovirus occurred in the 2009-10 season, with 207 symptomatic patients. This was followed by 126 cases in 2010-11, and 148 cases in 2011-12. In these first three years 43, 42 and 39 wards respectively were host to a patient that was infectious during the same hospital visit (not necessarily infectious when they visited the ward), with at least 14 of those wards confirmed to have hosted a patient when symptoms occurred.

Nationally, there was an early peak of cases in 2012-13, followed by typical prevalence throughout the remainder of the winter season (Adams 2013), but a different pattern was seen in OUH. There were only 22 symptomatic cases and 4 confirmed affected

wards - the lowest by far for all four seasons. Figure 4.4.4 and Figure 4.4.5 show the times of the hospital admission and discharge for each patient (i.e. the length of their entire hospital stay), coloured by the ward they were in when symptoms were first reported in the database.

Table 4.4.1: Summary of norovirus patients 2009-2013

	2009-10	2010-11	2011-12	2012-13
Number of patients				
Symptomatic	207	126	148	22
Sequenced	22	34	51	8
Number of wards affected				
Hosted infected patient[a]	43	42	39	12
Confirmed infected patient[b]	23	14	21	4
Number of ward stays[c]				
Symptomatic patients only	594	380	438	60
All patients	124,588	119,912	70,708	57,755
GII.4 variants sequenced				
New Orleans 2009	22 (100%)	34 (100%)	51 (100%)	-
Sydney 2012	-	-	-	4 (50%)
Recombinants	-	-	-	4 (50%)
Genetic clusters (cut off: 1 SNP)				
Total	15	21	36	7
Max. sequences per cluster	5	4	7	2
Genetic clusters (cut off: 3 SNPs)				
Total	12	8	13	3
Max. sequences per cluster	10	9	16	4
Genetic clusters (cut off: 10 SNPs)				
Total	6	8	8	2
Max. sequences per cluster	17	9	19	4

a. Wards which hosted a patient who were infected, had been, or who later became infected, during the same hospital visit. **b.** Wards where patient symptoms were first reported in the database. **c.** Number of patient ward stays (of any length) on a ward which hosted a patient who was infected, had been, or who later became infected, during the same hospital visit.

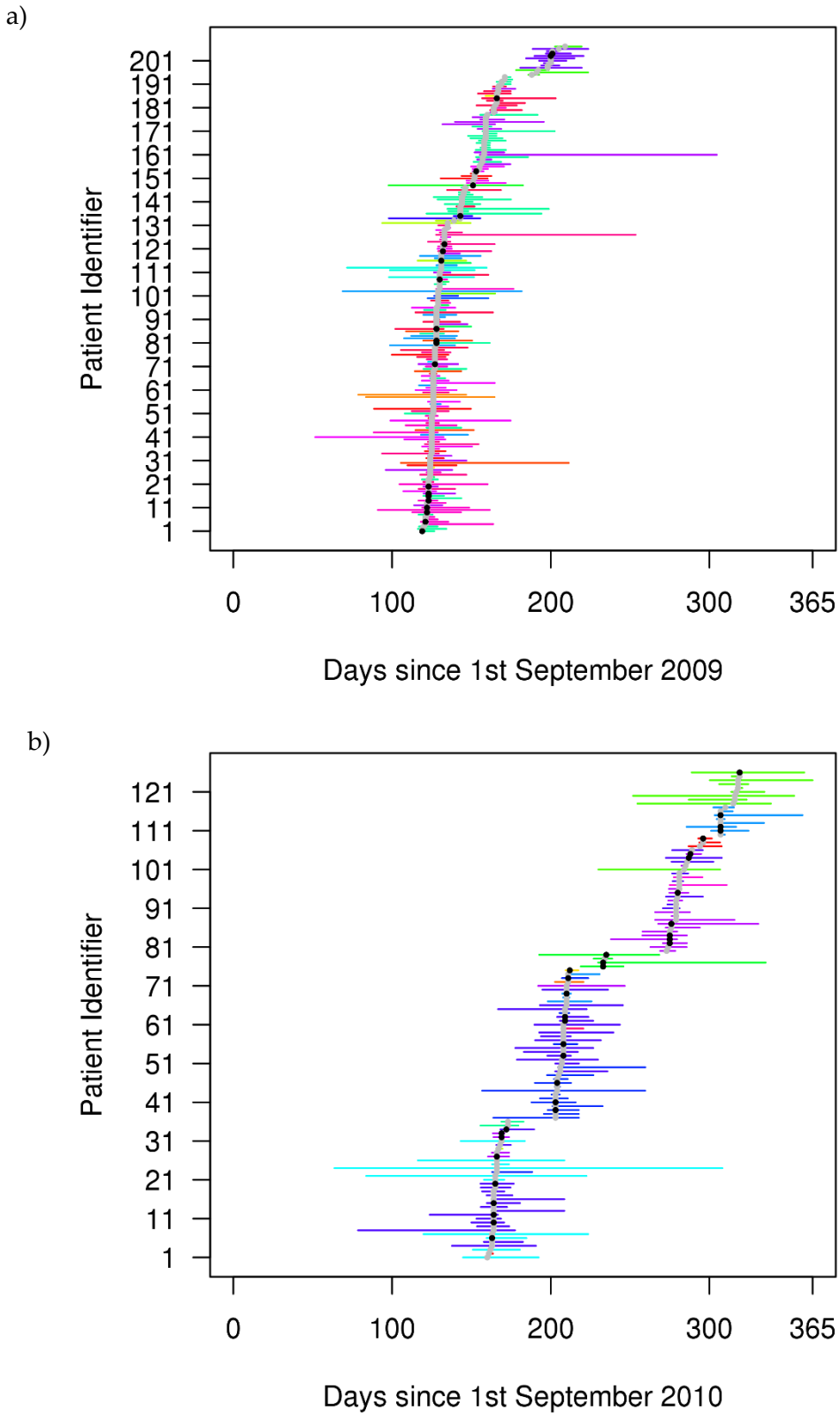
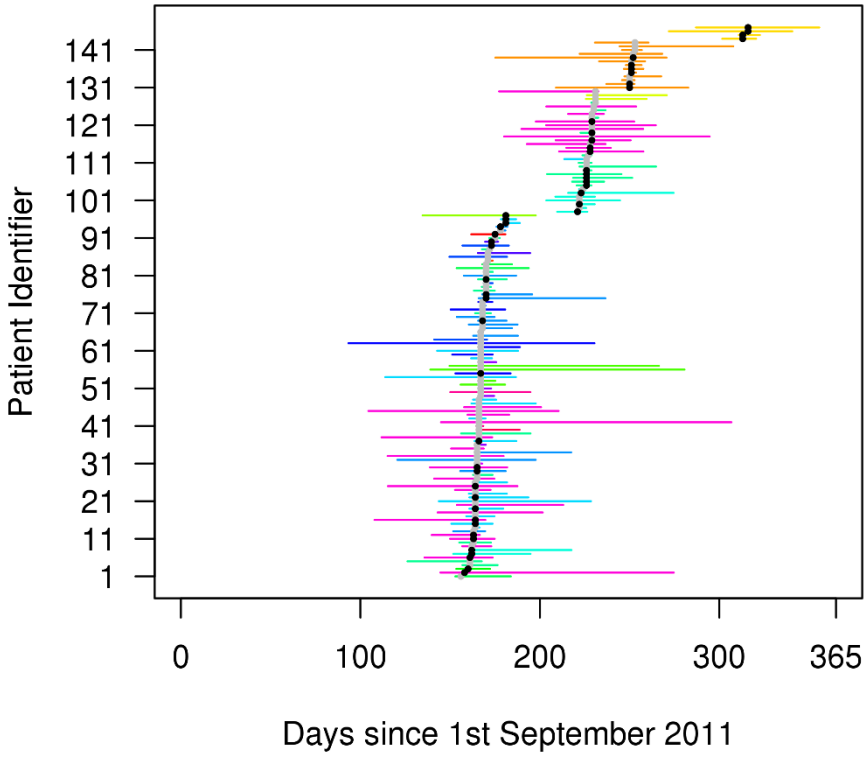


Figure 4.4.4: Hospital stays for symptomatic patients with norovirus for a) 2009-10 and b) 2010-11 seasons. Each horizontal line represents a hospital stay, defined as from the first day a patient was admitted to hospital until the day they were discharged. Circles denote the time at which symptoms were first reported - black if a sample was sequenced, and grey otherwise. Colours represent the ward on which they were staying when symptoms were reported in the database.

a)



b)

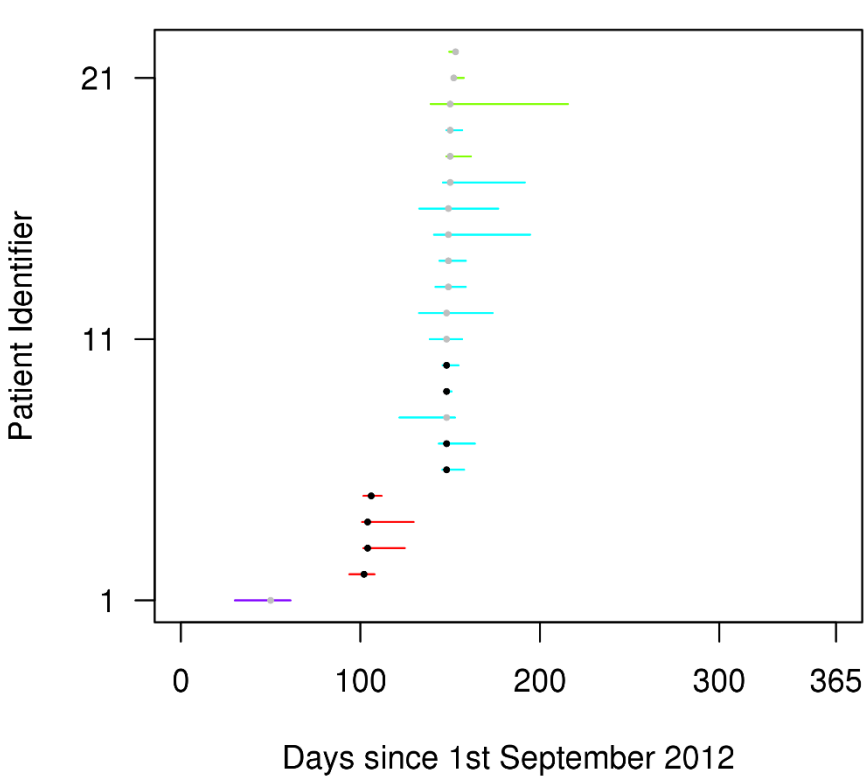


Figure 4.4.5: Hospital stays for symptomatic patients with norovirus for a) 2011-12 and b) 2012-13 seasons. Each horizontal line represents a hospital stay, defined as from the first day a patient was admitted to hospital until the day they were discharged. Circles denote the time at which symptoms were first reported - black if a sample was sequenced, and grey otherwise. Colours represent the ward on which they were staying when symptoms were reported in the database.

Obtaining samples for sequencing was difficult at the start of the study, so only 22 isolates were obtained and sequenced to high quality for 2009-10. However, the study gathered momentum, with 34 and 51 samples sequenced in 2010-11 and 2011-12. All of these isolates were typed to be of the New Orleans 2009 strain (Table 4.4.1). In 2012-13, eight samples were sequenced. Of these, half were found to be of the Sydney 2012 type, and the remaining four were recombinants with a New Orleans 2009 ORF1 and Sydney 2012 ORF2/ORF3, revealing a mixed strain epidemic.

The diversity of the genetic data for each season is captured in the form of a heat map in in Figure 4.4.6 and Figure 4.4.7. Each sequence along the x-axis is compared to those along the y-axis, and the colour of their intersecting square indicates the number of single nucleotide polymorphisms (SNPs). Pairs of sequences that are identical or have high similarity are coloured dark teal, whereas as pairs of sequences which are most diverse are coloured dark brown (1000 or more SNPs). In between these extremes, the palest cream represents around 570-600 nucleotide differences. The sequences are ordered according to similarity via hierarchal clustering, as shown by the dendrogram. These are not phylogenetic trees, but do give a view towards the relatedness of sequences as defined by the SNP differences. There is clearly one highly diverse sequence in the 2011-12 season, with a minimum of 963 and maximum of 1166 differences from the samples sequenced from other patients. In 2012-13, the brown and teal sections represent the dissimilarity between the four Sydney 2012 strains on the right, and the four recombinants on the left.

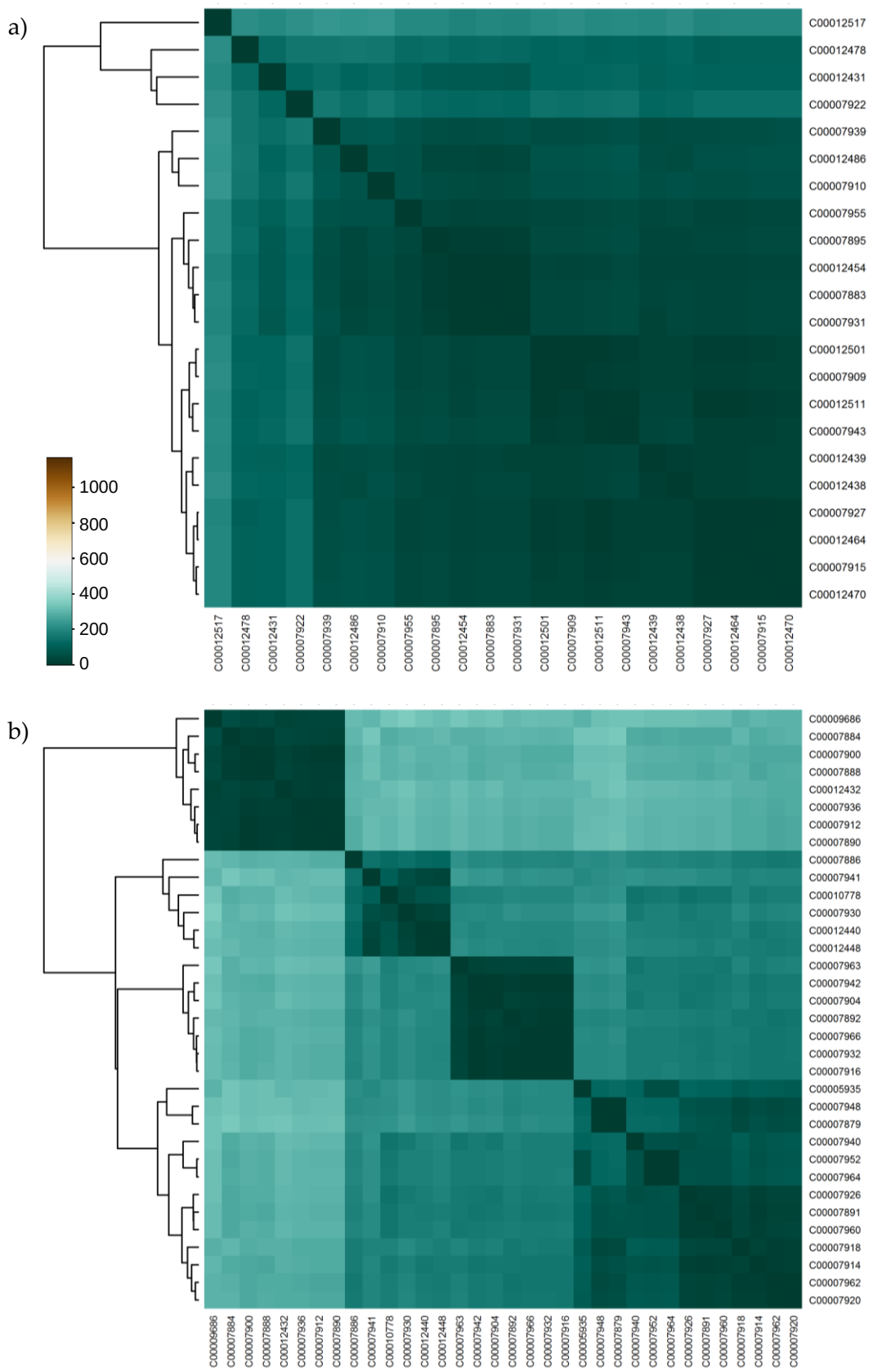


Figure 4.4.6: Heat map showing the pairwise number of SNP differences between patient sequences for the a) 2009-10 and b) 2010-11 seasons. Dark teal represents sequences that have the highest similarity, and dark brown those that have the greatest diversity (up to 1166 SNPs). The dendrogram gives the clustering of sequences by SNP distance.

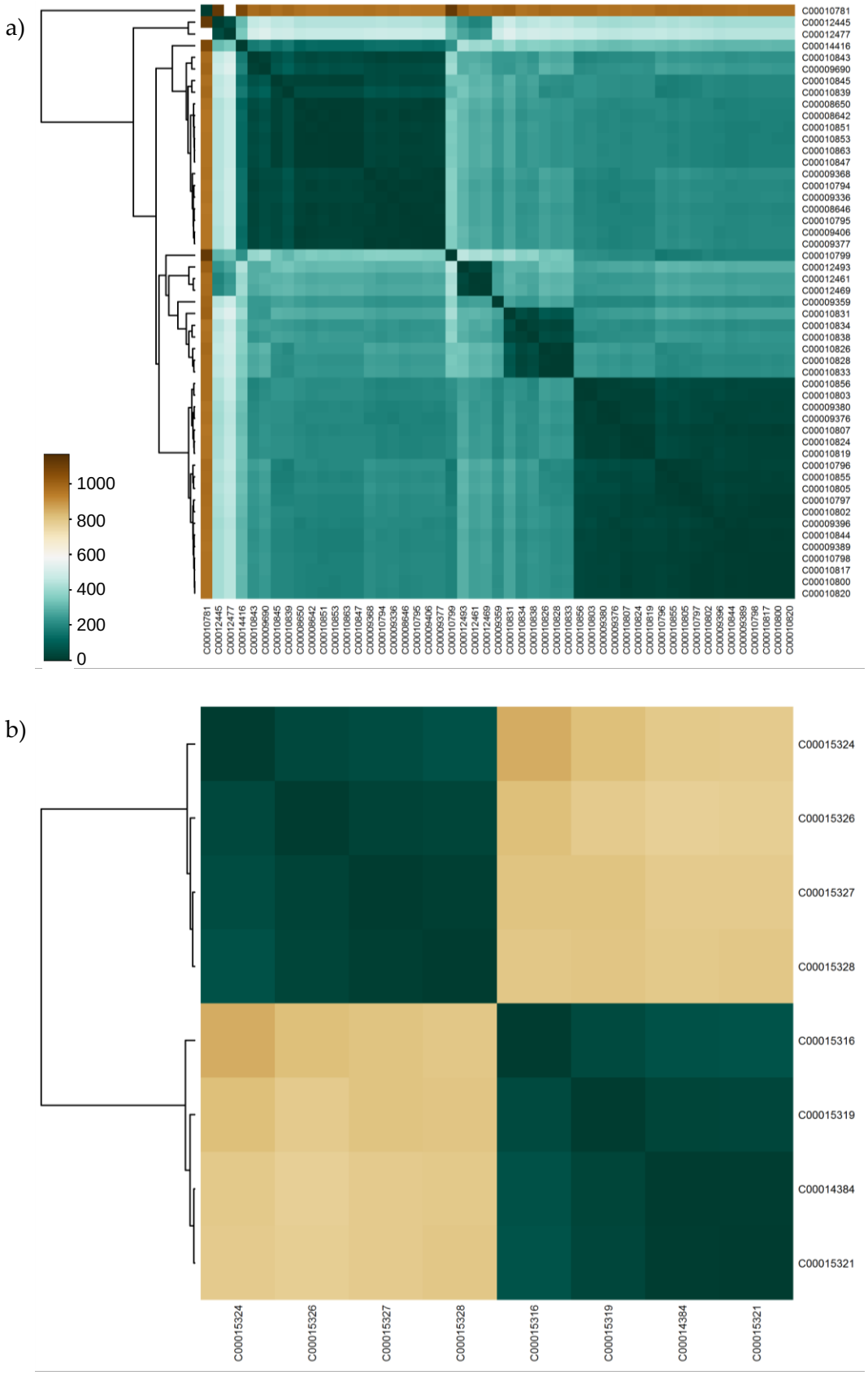


Figure 4.4.7: Heat map showing the pairwise number of SNP differences between patient sequences for the a) 2011-12 and b) 2012-13 seasons. Dark teal represents sequences that have the highest similarity, and dark brown those that have the greatest diversity (up to 1166 SNPs). The dendrogram gives the clustering of sequences by SNP distance.

I grouped patients with sequences together into clusters or 'STs' so that each isolate is a maximum number of from at least one other member of the same cluster. I investigated three different thresholds the SNP cut off: 1, 3, and 10 SNPs across the whole genome. Using the evolutionary rate of 5.29×10^{-3} substitutions per site per year from Section 4.3.1, these represent a time period of 9, 28 and 91days respectively. Norovirus is a self-limiting disease, with symptoms usually lasting around 3-5 days, so these thresholds should be conservative in allowing transmission between patients, despite the small possibility of sequencing error. The season with fewest genetic clusters was 2012-13, reflecting the fact that this season had the fewest sequences sampled (Table 4.4.1). In 2009, even at 10 SNP cut off, the 22 sequences were split across 6 clusters, five of which represented singletons that had no other sequence within 10 SNPs, showing that strain types were relatively divergent.

I ran the transmission model using all three SNP thresholds to investigate what effect, if any, the choice had on the results. Since, in general, the parameter estimates appeared to be robust to the choice of SNP threshold, here I only present the results of the most stringent (1 SNP) threshold. Results of the analyses under the 3 and 10 SNP thresholds are summarised in Appendix C.

4.4.4 Comparison of Norovirus Dynamics

Stochastic transmission model results revealed that transmission events traced to another known patient in the hospital represented the greatest risk of infection. For ease of notation, this will be referred to as 'patient-to-patient transmission', though it is

important to consider that such transmission events need not only to have occurred through actual contact, but also through a shared environment. The next greatest risk of infection came from the background within the hospital, followed by the background infection outside the hospital. The posterior distributions for the rate of infection via each of these routes are shown in Figure 4.4.8. The rate of patient-to-patient transmission, β_1, was estimated at up to 164 infections per 10,000 patient bed days (Table 4.4.2). Background infection within the hospital from unidentified sources was estimated to occur at a rate, β_0, up to 2.16 per 10,000 patient days in 2011-12. Background infection rates outside the hospital, β_1, accounted for less than 0.0005 infections per 10,000 patient days in all four seasons. This negligible rate of transmission from outside in the hospital was underlined by the low probability of colonisation on entry to the hospital, which was less than 0.001% in all four seasons.

Across seasons, the rate of patient-to-patient transmission varied from 65 infections per 10,000 patient bed days in 2012-13, to 165 in 2011-12. These rates differed significantly, as evidenced by the non-overlapping credible intervals in Table 4.4.2, suggesting that, as a susceptible patient on the same ward as someone infectious, the risk of being infected varied as much as 2.5-fold over the four year period. A different pattern was seen across seasons for the background rate of infection within the hospital, which ranged from 0.98 infections per 10,000 patient days in 2009-10 to 2.16 in 2011-12. Again, this shows evidence of a significant difference in rate, with the risk of infection increasing two-fold over the three seasons from winter 2009 to summer 2012.

Patient-to-patient transmission was modelled as occurring either directly, or from contamination remaining in the ward after the donor patient has been discharged. This distinction was made via the contamination multiplier, ψ. It was estimated that contamination remaining in the ward after the donor patient has been discharged accounted for 21-37% of the number of direct infections.

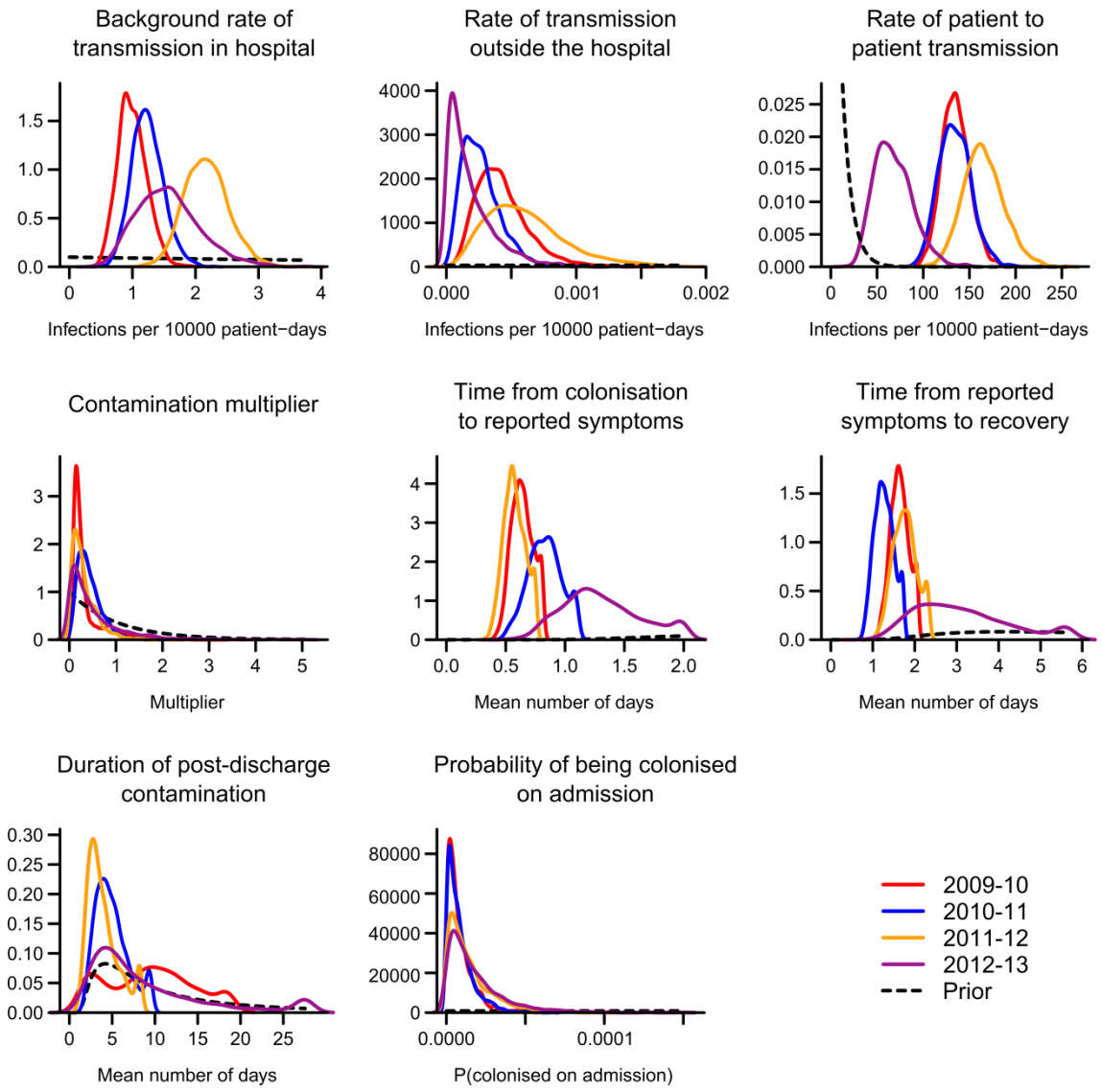


Figure 4.4.8: Comparison of parameter distributions for four seasons of the transmission model. Marginal posterior distributions for all parameters in the analysis; the prior distribution is shown with a dash black line, and the posteriors for the 2009-10, 2010-11, 2011-12 and 2012-13 seasons are in red, blue, orange and purple respectively.

Table 4.4.2: Summary of transmission parameter estimates by season.

Parameter	2009-10	2010-11	2011-12	2012-13
β_0 (infections per	0.98	1.22	2.16	1.54
10^4 patient days)	(0.59, 1.48)	(0.80, 1.77)	(1.52, 2.89)	(0.75, 2.70)
β_3 (infections per	0.04	0.02	0.05	0.01
10^6 patient days)	(0.01, 0.09)	(0.01, 0.06)	(0.02, 0.14)	(0.00, 0.07)
β_1 (infections per	134.39	133.85	163.89	65.38
10^4 patient days)	(107.82, 165.98)	(102.74, 170.76)	(126.27, 211.15)	(34.64, 114.41)
ψ	0.21	0.37	0.24	0.31
	(0.07, 1.53)	(0.08, 1.04)	(0.02, 1.03)	(0.01, 2.22)
λ_1	0.64	0.83	0.57	1.27
(days)	(0.48, 0.85)	(0.56, 1.15)	(0.41, 0.79)	(0.78, 2.23)
μ_1	1.64	1.26	1.76	2.95
(days)	(1.23, 2.16)	(0.87, 1.85)	(1.24, 2.43)	(1.47, 6.85)
μ_2	9.52	4.63	3.43	6.38
(days)	(1.53, 21.29)	(2.19, 11.13)	(1.56, 10.58)	(1.92, 45.03)
ϕ	0.06	0.01	0.01	0.01
$(\times 10^{-3})$	(0.00, 0.30)	(0.00, 0.03)	(0.00, 0.05)	(0.01, 0.06)

Parameters are defined in Table 4.3.1. **Inf. /10^{-4} pat. days** Infections per 10,000 patient days.

The duration of infection within a patient was broken down into the time from colonisation to symptoms being reported, and then until recovery. The duration of infection before symptoms were reported in the patient database was estimated to be between 0.57 days in 2011-12 to 1.27 days in 2012-13. This represents more than a two-fold increase from one season to another. The time from symptoms to recovery ranged from 1.26 days in 2010-11 to 2.95 days in 2012-13.

The duration of contamination was variable both within a season and across years, ranging from 3.43 days in 2010-11, to 9.52 days in 2009-10. The estimated duration of contamination in 2012-13 fell in the between of these estimates, but it was much more variable, with the 95% CI extending to 45 days.

4.4.5 Main Routes of Norovirus Transmission

I interpreted the results to identify the dominant routes of norovirus transmission. Across all the four seasons, the dominant route of infection inferred was that of a patient acquiring disease from another known infectious case in the hospital (Figure 4.4.9). The highest proportion of cases attributed to direct patient transmission was 76% in the 2009-10 season. This proportion falls successively in subsequent seasons to a low of 59% in 2012-13.

I split direct transmissions further into those that occurred before symptoms were reported, and those that took place after (Table 4.4.3). In 2009-10 and 2011-12, the majority of infection took place after symptoms had been reported. This fits with the biology of norovirus, since diarrhoea and vomiting cause a vast dispersal of virions.

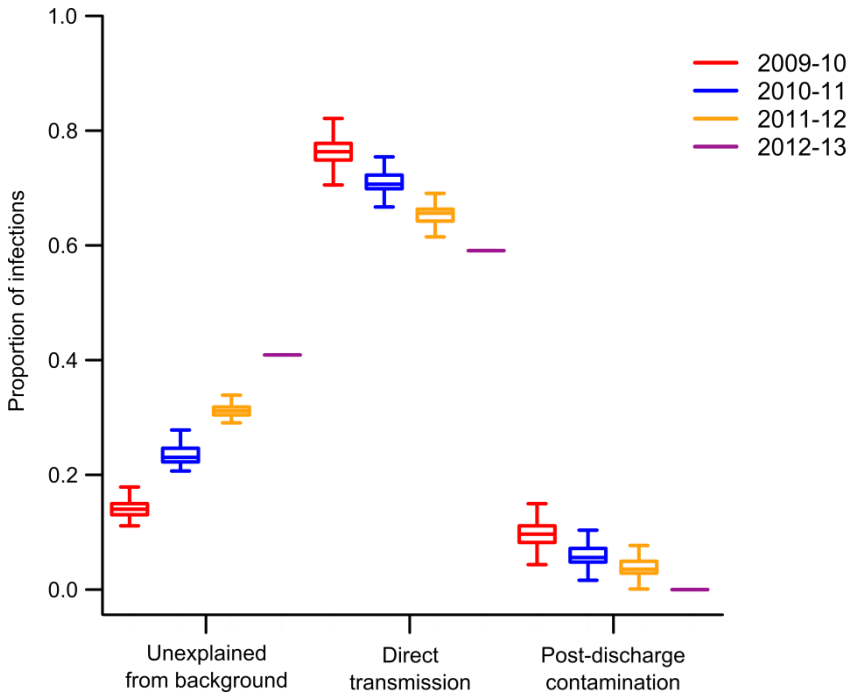


Figure 4.4.9: Sources of norovirus infection. Posterior distributions for the proportion of individuals infected from an unexplained source (either inside or outside the hospital), directly by another patient, or via contamination remaining after a patient has left a ward. Colours denote seasons: 2009-10 in red, 2010-11 in blue, 2011-12 in orange, and 2012-13 in purple.

However, infections attributed to a patient before symptoms are reported were inferred to account for up to 41% of transmission in 2010-11, compared to only 29% after they have been reported.

Ward contamination following discharge of infected patients did not appear to be an important route of transmission. Despite the high rate of transmission directly between patients, the proportion of infections due to the contamination remaining on a ward after a patient had been discharged was less than 10% in all four years.

Table 4.4.3: Posterior probabilities for the source of infection by season.

Source	2009-10	2010-11	2011-12	2012-13
Background:	**0.140**	**0.230**	**0.310**	**0.409**
	(0.115, 0.169)	**(0.214, 0.262)**	**(0.297, 0.331)**	**(0.364, 0.455)**
Inside hospital	0.121	0.214	0.290	0.409
	(0.096, 0.150)	(0.198, 0.246)	(0.276, 0.310)	(0.364, 0.455)
Outside hospital	0.019	0.016	0.021	-
	(0.019, 0.019)	(0.016, 0.016)	(0.021, 0.021)	
Direct:	**0.763**	**0.706**	**0.655**	**0.591**
	(0.715, 0.797)	**(0.675, 0.738)**	**(0.607, 0.683)**	**(0.545, 0.591)**
Pre-reported symptoms	0.314	0.413	0.310	0.318
	(0.251, 0.386)	(0.325, 0.500)	(0.255, 0.372)	(0.227, 0.455)
Post-reported symptoms	0.449	0.294	0.338	0.273
	(0.357, 0.517)	(0.206, 0.381)	(0.269, 0.400)	(0.136, 0.364)
Contamination:	**0.097**	**0.055**	**0.034**	**0.00**
	(0.053, 0.145)	**(0.024, 0.095)**	**(0.007, 0.083)**	**(0.00, 0.045)**
Pre-reported symptoms	0.024	0.047	0.007	0.00
	(0.005, 0.058)	(0.008, 0.079)	(0.00, 0.034)	(0.00, 0.045)
Post-reported symptoms	0.072	0.008	0.028	-
	(0.034, 0.116)	(0.000, 0.040)	(0.00, 0.069)	

4.5 Discussion

The aim of this chapter was to identify the dominant routes of transmission of norovirus in the hospital setting, and contextualise that within the wider evolution and epidemiology of GII.4 norovirus. I found that direct person-to-person contact was overwhelmingly the dominant route of infection for norovirus on hospitals, responsible for up to 76.3% (71.5, 79.7) of new infections in 2009-10 (Table 4.4.3). Almost all of this transmission occurred during overlapping ward stays, ruling out an important role for ward contamination as a vector of norovirus transmission ($PP < 0.145$ in all four seasons). This high rate of nosocomial transmission contrasts with recent work carried out by my colleagues in OUH, in which they discovered that 75% of *C. difficile* cases in Oxfordshire could not be attributed to direct transmission from other symptomatic cases in the hospital (Eyre et al., 2013; Walker et al., 2012; Cule and Donnelly, submitted).

As a susceptible patient, the risk of infection was up to 100 times higher when sharing a ward with someone infectious compared to the background rate elsewhere in the hospital (2009-10 and 2010-11 seasons, Table 4.4.2). This background rate includes unsampled sources of infection such as asymptomatic carriage, visitors, staff and contamination not attributed to a patient. Further, the rate of infection due to patients becoming infected outside of the hospital was negligible, estimated to be less than 1.4×10^{-7} for all seasons. This is an important result for infection management, as it suggests that independent introductions of norovirus to the hospital are generally rare; however, the presence of an infectious patient on a ward increases the risk of onward transmission hugely. The high proportion of cases attributed to a patient source is in

agreement with the results of Teunis et al. (2013), who showed 72% of cases in Netherlands had a direct transmission link using a transmission probability matrix. Gallimore et al. (2004b) found that 26% of staff and 33% of patients during an outbreak were asymptomatic, so this could explain how outbreaks of norovirus are initiated from within the hospital, rather than being brought in by symptomatic patients on admission.

I detected a different dynamic in the 2012-13 season compared to the other seasons, particularly the 2011-12 season preceding it. This was pertinent due to the emergence of the new Sydney strain of GII.4 norovirus in March 2012 and subsequent global spread (van Beek et al. 2013). The risk of infection for a susceptible patient from sharing a ward with an infectious patient was only 65.38 (34.64, 114.41) infections per 10,000 patients, the lowest by far of all four seasons covered in this study. This could have been due to the small sample size, but given sampling intensity increased as the seasons progressed, it more likely suggests that the 2012-13 strains were less transmissible, on average, than those in previous seasons, resulting in fewer secondary cases. This conclusion is supported by the high proportion of infections attributed to background pressures in the hospital compared to other seasons (PP = 40.9% (36.4, 45.5) in 2012-13, compared to PP = 31% (30, 33) in 2011-12, the season with the second highest proportion).

Interestingly, around 41.3% (32.5, 50.0) of infections in 2010-11 and 31.8% (22.7, 45.5) in 2012-13 were attributed to have come from patients before they were confirmed to be symptomatic. Norovirus is most easily spread through diarrhoea and vomiting (Caul 1994; Patel et al. 2009), so such a high prevalence of infection occurring before these

symptoms had been noted is unusual. However, it could reflect the lag between a patient experiencing symptoms, and nursing staff recording them in the patient database.

I assumed a single rate of recovery in this model, but there is evidence to suggest that shedding of norovirus can continue in some individuals for many weeks after symptoms have been resolved (Siebenga et al. 2008). If two patterns of infectiousness (symptomatic versus prolonged shedding) had occurred without having been explicitly accounted for, the posterior distribution for the time from test to recovery would have had a wide, possibly multi-modal, distribution. The 2012-13 season showed evidence for more variable recovery times than the other seasons, but the 95% credible interval was between one and seven days, and thus not consistent with long-term shedding patterns. Instead, I propose that the apparent slower recovery rate in 2012-13 was due to the prevalence of the Sydney 2012 strain. Four of the eight sequenced samples were typed to be this strain, and the remaining four were recombinants with a New Orleans ORF1 and Sydney 2012 ORF2/ORF3. A previous study in China by Mai et al. (2013) has suggested that symptoms are more severe when infected with the Sydney 2012 strain compared to previously circulating variants, and this might support a longer period of recovery. While the 2012-13 genomes were not exclusively Sydney 2012 variants, the presence of a dynamically different strain could explain the differences in recovery time and low secondary transmission rates.

One of the fundamental challenges with reconstructing transmission using whole genome sequences is interpreting genetic divergence in epidemiological terms. In this study, I defined strains or 'STs' in order to rule out transmission between patients infected by less closely related viruses. I defined the STs according to a SNP threshold, the choice of which could plausibly have affected the results. Choosing a threshold that was too stringent could have meant that isolates representing a transmission were not linked, prematurely ending a transmission chain and leading to an underestimate of patient-patient transmission (either directly or via a contaminated environment). On the other hand, an overly lenient threshold could have caused false inference of transmissions, and therefore led to an overestimate of the number of patient-patient transmissions, and an underestimate of infections coming from outside of the hospital. To test the robustness of inference under the model, I compared three SNP thresholds. Generally, there was good concordance in estimates (Appendix D), so I presented only the results from the 1 SNP threshold in this chapter. In 2010-11, 2011-12 and 2012-13, the 1 SNP threshold produced a slightly higher estimate of background rate of unexplained transmission compared to the 3 and 10 SNP thresholds, and in 2012, the 1 SNP threshold gave decreased estimates of direct transmission compared to the more relaxed thresholds. The posterior distributions in both scenarios overlapped, suggesting that the threshold choice did not have a substantial effect on parameter estimates. This shows the importance of choosing a SNP threshold relevant to the question. Here I was interested in the amount of direct patient transmission, and the most stringent 1 SNP threshold represented the most conservative estimate of this.

Splitting the study by season allowed dynamics to be compared across years. Most notably, the results alluded to differences in infection dynamics due to the strain type of GII.4 norovirus. In the first three seasons of the study period, all sequenced isolates were of the New Orleans 2009 strain type. However, no strains of this type were detected in 2012-13, when Sydney 2012 and New Orleans-Sydney recombinants took over (Table 4.4.1). This reiterates the importance of studying evolution and transmission side by side.

Previous estimates of the evolutionary rate have focused on only a limited region of the genome, and vary from 2 to 6 substitutions per genome per month (Bok et al. 2009; Bull et al. 2010; Siebenga et al. 2010). The rate calculated in this chapter fits at the low end of those estimates, which is to be expected given that I calculated a rate averaged over the whole genome, compared to previous estimates focusing on the diverse VP1 region or *in vitro* studies of the fallible RNA-dependant RNA polymerase. Most importantly, this updated rate gives a starting point for understanding norovirus transmission from a whole genome perspective.

Recombination is clearly important in the evolution of norovirus, with two recombinant sequences observed in a small sample of only 22 isolates from winter 2012-13, and a further four out of eight isolates in the transmission study from the same season. The inferred recombination breakpoint fits with other recent studies including Fonager et al. (2013), who found similar results to those presented here in Denmark, and a much larger study by Eden et al. (2013), who identified intragenotypic recombination events in

major GII.4 norovirus strains. An advantage of the clustering method here is that it does not make any assumptions regarding recombination, which would make relatedness testing via phylogenetic methods problematic, and therefore confers a degree of robustness to outbreak analyses even when emerging, possibly recombinant, strains are suspected to be involved.

Chapter 5: Zoonotic Transmission of *Campylobacter*

5.1 Campylobacter

Campylobacter is a genus of 18 species of bacteria, and is one of the most commonly identified causes of bacterial gastroenteritis – causing diarrhoeal disease up to seven times more frequently than other species including *Salmonella*, *Shigella* and *Escherichia coli* (Allos 2001). *Campylobacter* species are gram-negative, spiral-shaped bacteria that thrive in low oxygen, warm conditions – growing optimally in air with around 5% oxygen content (Park 2002). The cells range in size from 0.5 to 8.0μm long, and 0.2 to 0.5μm wide (Taylor 1992). They are motile due to the presence of either one or two flagella on the poles, which causes their movement to be corkscrew-like due to their spiral shape and appears to be suited to viscous environments (Ferrero and Lee 1988).

Campylobacter may have been discovered by Escherich in 1886, and was first cultured in the first part of the twentieth century after being recognised as a cause of abortions in sheep (Escherich 1886; McFadyean and Stockman 1913). However, it was not until the 1950s that the disease was associated with human gastroenteritis, and in 1963 the group of organisms were first separated from the *Vibrionaceae* group and the new genus named *Campylobacter* from the Greek for 'curved rod' (King 1957; King 1962; Sebald and Véron 1963). Despite this, *Campylobacter* has only fully been recognised as an important human pathogen in the last 30-40 years (Dekeyser et al. 1972). Of the campylobacter species that are pathogenic in humans, 90% is caused by *C. jejuni*, with the next most prevalent

cause of disease being *C. coli.* For the purposes of this section and the remainder of this thesis, the focus of discussion will remain on these two main species.

5.1.1 Epidemiology and Incidence

Campylobacter is a widespread zoonotic species, carried (often asymptomatically) in a range of mammalian and avian species. It is also regularly isolated from environmental sources such as marine waters and sewage, probably due to faecal contamination (Jones 2001; Lund et al. 2004; Bull et al. 2006b). The ideal temperature for growth appears to be 37-42℃ in host species, but in the environment the bacteria survive best in cooler, damp conditions (Hazeleger et al. 1998; Park 2002). In temperate conditions, there appears to be a peak in incidence in the summer, and possibly also in spring and autumn, though reasons for this pattern have not yet been fully established (Nylen et al. 2002; Louis et al. 2005; Meldrum et al. 2005).

The main risk factor for human *Campylobacter* infection in developed countries is handling and consuming poultry (Corry and Atabay 2001; Wingstrand et al. 2006). This is unsurprising, given that a study in South Wales found that 75% of sample of raw chicken from local supermarkets was contaminated with *Campylobacter* (Harrison et al. 2001), and a similar study in the US found 71% of retail raw chicken samples were contaminated (Zhao et al. 2001). In addition, it has been suggested there may be up to 500 infectious organisms in a single drop of chicken juice (Allos 2001), with up to $10^{3.91}$ *C. jejuni* cells recovered from a single chicken wing (Kinde et al. 1983). Human disease is also often associated with other farm species such cattle, sheep and pig (Wilson et al.

2008; Sheppard et al. 2009), and case-control studies have found that eating out at restaurants, red meat, pork, seafood, barbecues and unpasteurised milk are all potential risk factors (Kapperud et al. 2003; Neimann et al. 2003; Friedman et al. 2004). In contrast, the disease is hyper-endemic in developing countries. The prevalence in young children is particularly high, affecting up to 40-60% of children under the age of five (Calva et al. 1988). This may be attributable to weaning, as the child no longer receives antibodies from the mother's milk, and has not yet developed immunity of their own (Nachamkin et al. 1994; Coker et al. 2002). Asymptomatic infections are common in developing countries, as are those with multiple strains (often alongside other enteric pathogen species) despite these observations being rare in developed nations (Glass et al. 1983; Taylor et al. 1988; Taylor et al. 1993).

The true number of *Campylobacter* cases is unknown, as it relies on individuals both seeking medical attention and a sample being taken to confirm the diagnosis. In the US, it has been estimated that only 1 in 38 cases of infection are actually reported (Mead et al. 1999). In the UK, it is suggested that nearer to 50% of affected patients go to their general practitioner, but still less than half of those cases go on to be laboratory confirmed (Frost 2001). Figure 5.1.1 shows the number of laboratory reported cases in England and Wales from 2000 to 2011, plotted by month. The seasonal pattern is clear, with peaks occurring in the summer months. There is an upwards trend from 2005 onwards, which fits with patterns seen in other developed countries such as the US (Gilliss et al. 2013).

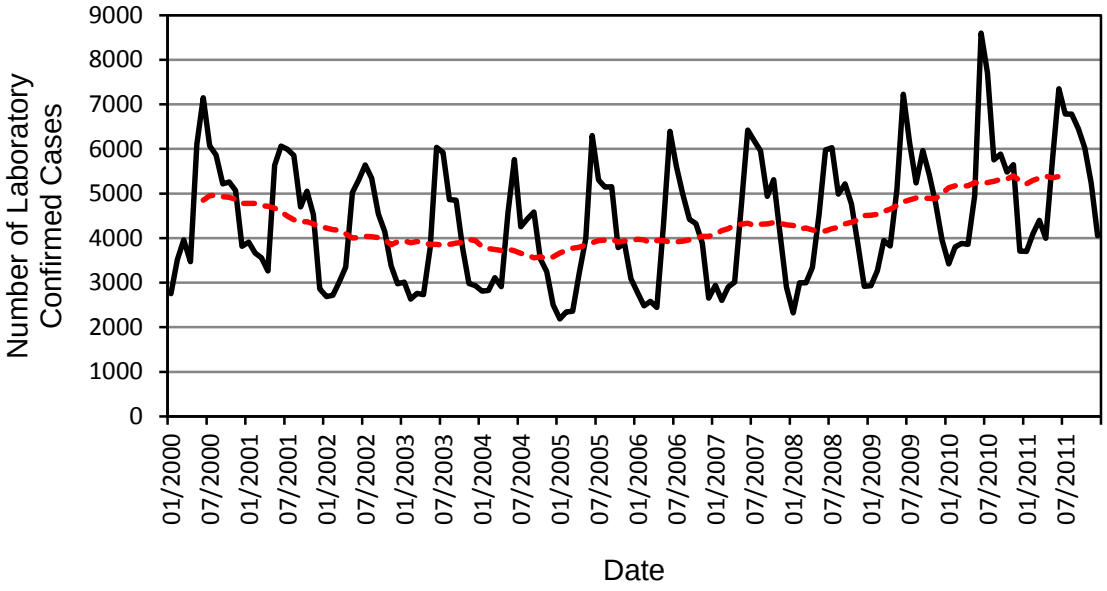


Figure 5.1.1: Reported cases of *Campylobacter* in England and Wales. The monthly number of *Campylobacter* cases (all species) reported to the Health Protection Agency from 2000 to 2011 (black), and 12 monthly moving average (red). Data from: Health Protection Agency (2012).

5.1.2 Characteristics of Human Infection

The infectious dose of *Campylobacter* is much higher than *Shigella* or *Giardia*, with 800 to 1,000,000 organisms required to cause infection in up to 50% of people (Black et al. 1988). Due to this, human to human transmission is rare (Allos 2001). Unlike norovirus, which has a low infectious dose, outbreaks of *Campylobacter* are rare in day care centres and health care institutions. Sporadic cases are thought to make up as many as 99% of cases, and outbreaks are usually due to a contaminated water or food source as opposed to a long chain of human transmission (Moore et al. 2005).

Infection with *Campylobacter* displays the usual symptoms of gastroenteritis within 2-5 days of colonisation: diarrhoea (containing blood or leukocytes), fever and abdominal cramps (Gillespie et al. 2006). In this sense, infection resembles that of other such bacterial pathogens including *Salmonella*, *Shigella* and *Yersinia*. Many patients also suffer

from fever, with temperatures over 40°C. Similarly to norovirus, treatment usually focuses on rehydration, though antibiotics may be given under certain circumstances, including if the patient is immunocompromised or suffering particularly bad symptoms (Blaser 1997). Identification of *Campylobacter* as the infective agent is done by culturing the organism from stool sample and PCR analysis (Stonnet and Guesdon 1993). Culturing is a slow process, requiring up to five days before a result may be produced complete with antibiotic sensitivity analysis, and detection of species may be limited according to the medium used (Moore et al. 2005).

Symptoms usually last for around a week, regardless of whether antibiotics have been given. However, infection has been known to last for several weeks in some cases, and bacterial shedding can persist several weeks after symptoms have cleared, unless antibiotics have been prescribed (Kapperud et al. 1992). Whilst most cases are self-limiting, there are possible secondary effects of infection as the result of the bacteria spreading further afield from the gut. Bacteraemia are generally rare, with a low case-fatality rate (estimated to be 0.05 per 1000 infections), and are of greatest risk in patients such as the young, elderly and immunocompromised (Skirrow et al. 1993). Infection with *Campylobacter* can also lead to more serious complications including Guillain-Barré syndrome (Nachamkin et al. 1998), Miller Fisher syndrome (Taboada et al. 2007), and reactive arthritis (Pope et al. 2007).

Until the late 1990s, fluoroquinolines and macrolides were the preferred antibiotic treatment, as they were also able to treat other pathogens with similar symptoms such

as *Salmonella* and *Shigella* in the time before a diagnosis was laboratory confirmed. However, strains of *Campylobacter* resistant to both of these treatments have been discovered, and the increase in their prevalence is synonymous with overuse in human and veterinary prescribing, particularly of food animals (Sam et al. 1999; Piddock et al. 2000; Allos 2001; Engberg et al. 2001). Resistance to erythromycin continues to be low, around 5%, despite continued medical and veterinary use (Allos 2001; Wimalarathna et al. 2013). Whilst this is good news, the limited number of treatments available for both humans and livestock is of great concern (Moore et al. 2006).

5.1.3 Evolution

The *C. jejuni* reference genome is around 1.6Mb in length, which is relatively small compared to *E. coli*, which has a genome of ~4.6Mb. This limited genome size may explain the lack of ability to obtain energy from carbohydrates (either through fermentation or oxidization), and the resulting difficulties in finding a suitable culturing medium for growing *Campylobacter* in the laboratory (Griffiths and Park 1990). The small genome also means that *Campylobacter* does not have the DNA repair mechanisms found in many larger bacterial genomes, which may contribute to fallible replication and rapid variation in sequences, possibly resulting in diversity within a host resembling a quasispecies (Parkhill et al. 2000). Parkhill et al. (2000) also noticed variable regions consisting of homopoly (repetitive single nucleotide sequences), which they noted may be important in survival.

The highly diverse genome of *Campylobacter* has proved problematic in understanding how it evolves, even for the 'core' genome (the genes present in all strains). Wilson et al. (2009) used a combined population genetics and phylogenetic approach to show that *C. jejuni* evolves rapidly, though purifying selection removed up to 60% of the resulting mutants, yielding an estimated mutation rate of 3.23×10^{-2} mutations per kilobase per year. However they suggest that despite this slow rate, the large effective population size means that novel mutations can occur at any site in the population after only a week. In addition, recombination has been found to be very important in maintaining the genomic diversity in *Campylobacter*, occurring both within a species, and between them – particularly between *C. jejuni* and *C. coli* (Dingle et al. 2005; Fearnhead et al. 2005; Sheppard et al. 2008). Indeed, recombination has been estimated to generate diversity at twice the rate of mutation (Wilson et al. 2009).

5.1.4 Multilocus Sequence Typing and Genomic Sequencing

Multilocus sequence typing (MLST) for *C. jejuni* and *C. coli* is based on seven housekeeping genes: *aspA*, *glnA*, *gltA*, *glyA*, *pgm*, *tkt*, and *uncA* (Dingle et al. 2001; Dingle et al. 2005). These loci are 402 to 507 base pairs in length, and are spread around the reference sequence with a minimum gap size between positions of 70kb. The seven loci have been chosen such that there is sufficient diversity to be useful in typing, both specific and sensitive enough to discriminate between isolates, but also not influenced by positive selection or recombination which would make the resulting data less amenable to population genetic methods.

At each locus, the alleles are assigned an arbitrary numeric label, and thus each isolate has a profile of seven numbers. Each unique allelic profile is known as a sequence type (ST), and thus represents the full resolution of MLST. Clonal complexes represent groups of closely related sequence types, differing in alleles at three or fewer of the loci. Thus these complexes ideally give groups of sequences that have a more recent common ancestor with other members of the group compared to those strains outside it (Smith et al. 1993; Maiden et al. 1998).

The allocation of sequence types enabled work into host association and sources of human disease, and the genetic evolution within and across *Campylobacter* species (Wilson et al. 2008; Sheppard et al. 2009; Wilson et al. 2009). However, understanding ancestral relationships with MLST is not necessarily straightforward in a complex species such as *Campylobacter*. The sharing of alleles in clonal complexes is likely to imply recent shared ancestry, but the effect of recombination means that alleles at different loci may not have been inherited via the same route.

Campylobacter was first whole genome sequenced in 2000 (Parkhill et al. 2000). To date, the majority of studies using whole genome data have been limited to the announcement of one or two new sequences, and comparison to those sequences already publically available (Fouts et al. 2005; Pearson et al. 2007; Poly et al. 2007; Takamiya et al. 2011; Jerome et al. 2012; Wu et al. 2013). There have been three main studies with larger datasets. Lefébure et al. (2010) analysed 43 strains of *C. jejuni* and 42 strains of *C. coli* to investigate the species-specific core genes and identify a finite pan-

genome of similar size in both *C. jejuni* and *C. coli.* Sheppard et al. (2013a) analysed 26 new strains of *C. coli* to demonstrate the introgression of *C. jejuni* genes into *C. coli.* Both of the aforementioned datasets were combined with a further 80 genomes in an association study by Sheppard et al. (2013b), identifying genes involved with vitamin B_5 biosynthesis as frequently present in cattle, but absent in chicken.

5.2 Motivation

Many new and emerging human pathogens are zoonotic, making a crucial leap from being well established in animal populations to being able to transmit to humans. Viruses that have successfully achieved this and gone on to cause serious epidemics in humans include severe acute respiratory syndrome (SARS), Ebola fever, and various strains of influenza (Parrish et al. 2008). However, zoonosis is not limited to viruses, and there are a number of well-known bacterial zoonoses including *E. coli* O157:H7, *Salmonella* enteritidis, Plague (*Yersina pestis*), brucellosis and bovine tuberculosis (Corbel 1997; Caprioli et al. 2005; Duplantier et al. 2005; Velge et al. 2005; de la Rua-Domenech 2006). Parrish et al. (2008) describe three stages that take place during the emergence of new zoonoses. Firstly, a few pathogens gain the ability to be able to transfer into the new host, but with little or no further transmission. Then, occasionally, some of these will be transmit further throughout the host population in localised outbreaks, and finally, this extends into sustained transmission between hosts.

The chance of zoonosis is dependent on many factors in both the pathogen and the new host (see, for example, the reviews in Parrish et al. (2008) and Blancou et al. (2005)), but

of clear importance is the role of a change in interaction between original and new host, whether demographic, social or behavioural (Garnett and Holmes 1996; Dobson and Carper 1996b). In emerging human zoonoses, this interface can be either through direct (wildlife, pets) or indirect (foodborne) contact (Cleaveland et al. 2001; Daszak et al. 2001; Blancou et al. 2005). The capability of colonising more than one species clearly increases the scope of transmission and therefore survival of the pathogen, but it also relies on being able to rapidly adapt to the new host. Yet, this adaptation can come at a cost; research has shown that adaptation to one environment may cause reduced fitness in others (Giraud et al. 2001).

The genus *Campylobacter* comprises several widely distributed species of zoonotic pathogens, carried, often asymptomatically, in the gut microbiota of a range of mammalian and avian species, including, notably, those found on farms. As a consequence, opportunities for contamination of products including meat, poultry and unpasteurised milk with *Campylobacter* species can take place at any time from farmyard to consumption (Neimann et al. 2003). Traditional epidemiological analysis has been used to investigate the major sources of *Campylobacter* species infection in humans, but this has proven difficult because of unusual transmission dynamics. In contrast to other common gastroenteric pathogens such as *E. coli*, *Salmonella* and norovirus, *Campylobacter* infection in humans is sporadic, and rarely manifests as large outbreaks where a single infective source can be identified with confidence (Pebody et al. 1997). When large outbreaks of *Campylobacter* have occurred, they are usually associated with contaminated drinking water (for example, as described in Clark et al. (2003), Engberg et

al. (1998) and Palmer et al. (1983), amongst others) and there has been very little evidence towards direct human-to-human transmission (Allos 2001). Therefore, attention has focused on attributing human cases – at the level of population – to various reservoir populations, including animals farmed for meat and poultry, wild birds and mammals, and the environment (Wilson et al. 2008).

Genetic analysis through multilocus sequence typing (MLST) has proven useful in understanding the source of human infection, exploiting genetic differences that exist between *Campylobacter* strains that live in different reservoirs (McCarthy et al. 2007). Replicated in independent studies in England, Scotland and New Zealand, this approach has revealed that over 95% of human cases can be attributed to meat sources (chicken, cattle, sheep and pig), with 56-76% of cases attributed to poultry alone (Wilson et al. 2008; Mullner et al. 2009; Sheppard et al. 2009). While at the population level, MLST has the power to reveal the dominant sources of *Campylobacter* infectious to humans, at the level of individual cases there often remains considerable uncertainty as to the source population. Although some sequence types (STs) are statistically associated with a specific source - for example, some wild bird species are colonised with phylogenetically distinct *Campylobacter* lineages (Sheppard et al. 2011; Griekspoor et al. 2013), and ST-257 and ST-61 are associated with chicken and ruminants respectively (Sheppard et al. 2011) - others appear associated with numerous host species. Indeed, some of the most common lineages associated with human disease including ST-21, ST-45 and ST-828 are found frequently in both chicken and cattle. This has led some to argue that STs of *Campylobacter* differ in their level of host specificity,

with some designated as specialists and others generalists, for example as in Gripp et al. (2011).

With the advent of whole genome sequencing, there is new hope of detecting fine-scale genetic structure below the level of ST (Enright and Spratt 2011; Wilson 2012). This could reveal previously undetected associations between sub-strains and particular hosts, potentially revealing that STs such as ST-21, ST-45 and ST-828 are not in fact genuine generalists, but are merely an aggregate of host-restricted sub-strains that appear to be generalists due to the limited resolution of MLST (Schouls et al. 2003). The improved resolution afforded by whole genome sequencing would then lead to more accurate source attribution, and provide further evidence to inform public health strategies and subsequently help reduce the current upwards trend in cases (Figure 5.1.1) as occurred after MLST analysis implicated poultry as the main source of infection in New Zealand (Sears et al. 2011). Yet, there is another possibility: STs such as ST-21, ST-45 and ST-838 commonly isolated from more than one host reservoir may actually be genuine generalists, capable of living on and transmitting between multiple host species or environments. In this scenario, whole genome sequences should reveal little or no clustering with reservoir even below the level of ST; instead we would conclude that these strains are indeed able to colonise hosts with a broad spectrum of physiological traits, including different immune systems, digestive tracts and body temperatures (Palmer et al. 2005).

5.2.1 Chapter Aims

In this chapter, I use whole genome sequencing to test the hypothesis that isolates from ST-21, ST-45 and ST-828, three of the most common human disease-causing sequence types of *C. jejuni* and *C. coli*, are in reality strongly host-restricted below the level of sequence type, and whether it is possible to use this information to infer the source population of human isolates more accurately. This analysis provides additional insights into the rate of zoonotic migration of these sequence types between source populations.

5.3 Methods

5.3.1 Overview

To test the hypothesis of sub-ST host restriction in *Campylobacter*, I will estimate the number and rate of zoonotic transmissions in the ancestral history of isolates sampled from various sources including chicken, cattle, pig and wild birds. If strains are perfectly host restricted below the level of ST, we would expect to see isolates sampled from the same source reservoir clustering into one or a small number of distinct clades within the phylogeny. Under the general hypothesis, *Campylobacter* isolates from the same source population will be scattered throughout the tree, interspersed with isolates from different source populations, and there will be no association between source and evolutionary lineage.

Central to the analysis is the estimation of the number of migration events between source populations, which will be low when isolates are host restricted. At one extreme,

only a single transmission event is required for the initial founding of each source population. At the other extreme, if lineages are transmitting freely between source populations, there will be numerous transmissions between source populations. Formally, I will adapt the phylogeography approach of Lemey et al. (2009), treating source populations as if they were distinct geographic entities. The phylogeography approach, which is implemented in BEAST (Drummond et al. 2012) treats the population to which an ancestral lineage belongs as a discrete trait that evolves along the phylogeny according to the migration rate matrix.

5.3.2 Isolate Collections and Whole Genome Sequencing

Three sequence types were chosen as the focus of the study, ST-21, ST-45 of *C. jejuni*, and ST-828 of *C. coli*, because they are among the most common lineages that cause disease in humans, but unambiguous attribution of a source population has proven difficult using MLST data (Wilson et al. 2008; Sheppard et al. 2009). Isolates from these three STs were chosen from MLST collections (www.pubmlst.org/campylobacter, Jolley and Maiden (2010)), sequenced and assembled according to the protocol described in Sheppard et al. (2013b), and their exact isolate numbers are given in Appendix E. In total, 30, 28 and 42 sequences were obtained for ST-21, ST-45 and ST-828, respectively (Table 5.3.1). Chicken and cattle are the reservoirs most associated with human infection, so in all three STs these were sampled at with highest frequency. Since wild bird isolates were available for ST-45 and pig isolates for ST-828 these were also included, though they are known to be much less likely to cause human infection. MLST

shows there is little evidence for there being an environmental source of infection, so no environmental samples were taken.

Table 5.3.1: Source populations of isolates by ST.

Host Species	ST-21	ST-45	ST-828
Chicken	13	13	13
Cattle	7	9	10
Pig	-	-	6
Bird	-	3	-
Human	10	3	13
Total	30	28	42

After Illumina sequencing, the high coverage short reads of 25-50bp in length were *de novo* assembled using Velvet (Zerbino and Birney 2008), and the resulting contiguous sequences ('contigs') stored using BIGSdb (Jolley and Maiden 2010). These contigs were then compared to the NCTC11168 reference sequence (Genbank accession number: AL111168) to identify genes using a BLAST search. A reference sequence is a representation of the genes found in a species, often used as a baseline for comparison of newly sequenced genomes. NCTC11168 was the first *Campylobacter* genome to be sequenced and annotated in 2000 (Parkhill et al. 2000). It was subsequently updated and re-annotated in 2007, making it an important resource for locating and aligning *Campylobacter* genes (Gundogdu et al. 2007). Orthologous genes to the reference sequence were defined as homologous genes that had 70% or greater nucleotide identity, and less than 50% difference in alignment length. Genes for all isolates were then aligned using MUSCLE (Edgar 2004), and concatenated into a single sequence per sample, including gaps for missing genes relative to the reference. The resulting fasta

file of these sequences was the start point of my analysis, along with information on the host species from which the isolate was sampled.

5.3.3 BEAST Analysis

The analysis was implemented in BEAST v1.7.5 (Drummond et al. 2012), using a zoonotic transmission reconstruction model to infer the source of the infection (host species) along the branches (see Section 5.3.5 for further details). Given that human to human transmission is rare (Allos 2001), an ambiguity code was set up in the zoonotic model to allow the human isolates to have an 'unknown' source population. For each human isolate, the other host species in the analysis were given equal prior probability and thus most likely source of the human isolates could be inferred. This is edited in to the BEAST xml file by adding an `<ambiguity>` to the `<generalDataType>` as in the example below, where state 9 can be inferred to be either of states 1 or 2:

```
<generalDataType id="host.dataType">
     <!-- Number Of States = 2 -->
     <state code="1"/>
     <state code="2"/>
     <ambiguity code="9" states="12"/>
</generalDataType>
```

A constant population was assumed, due to computational efficiency compared to more complicated models of demographic change in what is already a complex model. All parameters were scaled in terms of the effective population size, N_e, by setting it to 1.0. The population scaled clock rate was $\theta/2$. For the zoonotic transmission reconstruction model, a uniform prior from 0.0 to 100.0 was assumed for the host migration rate, a gamma prior with mean 1.0 and scale 1.0 for the relative rates of migration, and a uniform (Dirichlet) prior for the host population equilibrium frequencies. The HKY85

model of nucleotide substitution (Hasegawa et al. 1985) was used with an uncorrelated log-normal relaxed clock and gamma rate heterogeneity with four categories (Drummond et al. 2006). A log-normal prior with a mean of 1.0 and standard deviation of 1.25 on the logarithmic-scale was assumed for the transition:tranversion ratio κ, and an exponential prior with mean 1.0 was utilized for the gamma shape parameter α. For the log-normal relaxed clock parameters, a uniform prior between 0.0 and 10.0 was assumed for the mean, and an exponential with mean 1.0 for the standard deviation. A uniform (Dirichlet) prior was used for the nucleotide frequencies. To calculate the number of discrete zoonotic transmissions across the branches, Markov jumping was performed using BEAST (Minin and Suchard 2008a; Minin and Suchard 2008b; Talbi et al. 2010).

The MCMC was run for 500 million iterations, with samples taken every 5000 iterations after discarding the first 10% of iterations as burn in. For each ST, the analysis was repeated four times with different initial values to check convergence and mixing, and these runs were combined for the final results. Unless otherwise stated, the posterior median was used for point estimates and the (2.5%, 97.5%) quantiles for credible intervals. The inferred ancestral host type for each branch in the phylogeny was taken to be the one with the highest posterior probability, and the trees were visualised using FigTree v1.4 (Rambaut and Drummond 2009). Due to a lack of sample date information, the analysis is given in coalescent time (denoted τ). However, this is calibrated in to years using an independent estimate of the mutation rate in *Campylobacter* of 3.23×10^{-5} substitutions per site per year from Wilson et al. (2009).

5.3.4 Accounting for Ancestral Recombination

There is much evidence to suggest that novel diversity in *Campylobacter* is generated frequently by the continued movement of genes between lineages more so even than by the evolution of new variants through mutation (Wilson et al. 2009; Sheppard et al. 2010). A high level of recombination leads to mosaic genomes with differing ancestral histories, and the full evolution of samples cannot be represented by a coalescent geneology. Instead, relationships can be represented using the more complex Ancestral Recombination Graph (ARG) or phylogenetic network (Hudson and Kaplan 1988; Griffiths and Marjoram 1996; Huson and Bryant 2006). Software such as SMARTIE can infer the most likely ARG of a sample, and to count to number of recombination events, but such implementations are limited to a small number of sequences (Bloomquist and Suchard 2010). Preferably, an analysis would be performed such as the one implemented by ClonalFrame (Didelot and Falush 2007), which identifies the clonal history of isolates in the sample whilst also estimating when and where recombination events took place. However, there are a number of limitations with such programs, not least being that they offer limited models of demography (such as not implementing the phylogeography model for understanding migration between hosts), have a lack of robust methods for comparison (Arenas et al. 2008), and involve a not-inconsiderable amount of computational time with whole genome sequence data. Thus, there are advantages in favour of using a phylogenetic tree building approach.

During preliminary analyses using BEAST, a relaxed clock model and gamma site heterogeneity were used to try and account for the effect of recombination. Informally,

this assumes that a recombination event resulting in high sequence diversity on a branch is equivalent to that branch having a relatively high mutation rate when compared to branches with no or low diversity recombination events, and is a step towards the model underlying ClonalFrame with the model flexibility of BEAST. However, these analyses diagnosed difficulties in the mixing of MCMC algorithm underlying BEAST, even notwithstanding the ClonalFrame-like model, with multiple runs of the same analysis frequently converging to different topologies. To overcome this, and account for the effect that recombination has in skewing the branch lengths of the dominant phylogenetic tree (Schierup and Hein 2000), homoplasious sites incompatible with the maximum likelihood tree were identified and removed, as well as sites with more than two alleles (Pupko et al. 2000; Guindon et al. 2010).

For the final dataset, only the biallelic sites compatible with the inferred phylogeny were included in the alignment, alongside all of the non-variable sites. The removal of homoplasies in this way resolved the mixing issues with BEAST.

5.3.5 Zoonotic Transmission Reconstruction

Recently, Lemey et al. (2009) have developed a model for phylogeographic inference and implemented it in BEAST. This method has been used in a number of geographic applications (Carnieli et al. 2011; Véras et al. 2011; Zehender et al. 2011; Lycett et al. 2012a; McAdam et al. 2012), but it can be used for any discrete trait or phenotype, such as gene reassortment between subtypes (Lycett et al. 2012b; Ward et al. 2013) and the identification of important ancestral hosts (Haß et al. 2011). Here, I used it to represent a

migration model between different reservoirs of *Campylobacter* defined by host species (Figure 5.3.1). Using the method, it is possible to report the posterior probability of any node in the tree being a particular state (for example, to infer the trait of the MRCA), and also to obtain relative rates of transition between states, for example to see if there is an association with certain traits being more likely to switch. In this chapter, the method is used to reconstruct the number of zoonotic transmission events of *Campylobacter* throughout the tree, and to obtain estimates for rates of migration between host species.

Lemey et al. (2009) model transitions between the discrete states as a Markov process, and thus the discrete traits model is analogous to the setup of a standard nucleotide substitution model as described in Section 2.6.1. A matrix $\boldsymbol{\Lambda}$ (equivalent to the $\boldsymbol{Q}$ matrix) is constructed from three main components: μ, the instantaneous transition rate, which acts to scale the trait transition rates into the units used for the tree; a matrix $\boldsymbol{S}$, which determines the rates of transition between each pair of states, relative to every other pair of states; and a diagonal matrix $\boldsymbol{\Pi}$, which contains the equilibrium frequencies of the traits of the time reversible model. The finite-time transition probabilities are then obtained by exponentiating $\boldsymbol{\Lambda} = \mu \boldsymbol{S} \boldsymbol{\Pi}$ (cf. Equation (2.12) in Section 2.6.1).

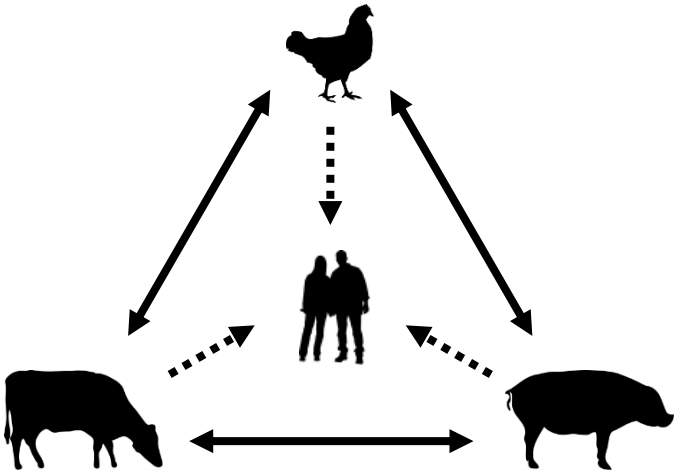


Figure 5.3.1: Zoonotic model of transmission. A schematic of the zoonotic transmission model implemented in BEAST, analogous to the nucleotide substitutions in Figure 2.6.1. Solid arrows represent the rates of transmission between hosts, Sij. Dashed lines represent transmissions into humans – these are modelled as ambiguous states at the tips of the tree, and are inferred in each iteration of the analysis and summarised at the end to give a posterior probability.

5.3.6 Markov Jump Counting

The discrete traits model can be used to identify the most likely trait that would have been observed for a branch in the tree, but it does not by default output what proportion of the branch length has been spent as that trait, nor how many swaps there have been to and from other traits (source populations) in that time. In other words, the evolutionary path of the discrete trait is hidden. Minin and Suchard (2008a) showed that it is possible to find an analytical solution for the moments of the counting process of discrete traits, allowing the number of transitions (or 'Markov jumps') across a branch to be recorded. This method was extended to a whole tree in Minin and Suchard (2008b), alongside tracking when the Markov jumps occurred. This latter extension allows the length of time spent in each trait along a branch to be calculated (known as the 'Markov rewards').

The method of Minin and Suchard (2008b) is simulation-free, meaning that it is much more computationally efficient than alternative, simulation-based methods such as the stochastic mapping of Nielson (2002), where histories of characteristics have to be repeatedly simulated and discarded if not compatible with the data. Talbi et al. (2010) implemented Markov jump counting in BEAST, where it can be run in conjunction with the discrete traits phylogeographic method of Lemey et al. (2009).

5.4 Results

5.4.1 Evidence for Ancestral Recombination

As noted, I included three STs of *Campylobacter* in the analysis, two of *C. jejuni* (ST-21 and ST-45) and one of *C. coli* (ST-828). Samples from chicken and cattle were obtained for all three STs, as previous work using MLST has shown these are the most common reservoir species for human infection. Wild bird isolates were also collected for ST-45, and pigs for ST-828. It is known that, even within an ST, there is evidence for considerable recombination detected using whole genome sequencing in *C. jejuni* and *C. coli* (Biggs et al. 2011). Recombination is problematic for phylogenetic analyses as it skews branch lengths in the tree, and leads to an overestimate in the substitution rate heterogeneity and loss of molecular clock (Schierup and Hein 2000). Therefore, I required that only polymorphic sites in the genome that represented the dominant underlying phylogeny should be used for the full analysis. This left sequence alignments with 51.7% of the original polymorphism in ST-21, 50.3% in ST-45, and 12.6% in ST-828 (Table 5.4.1). Whilst this clearly removes a sizable fraction of the polymorphic sites, a virtue of full-length genome sequencing is that the remaining sites are highly

informative about the underlying ancestral tree. MLST covers less than 0.5% of the genome with up to 21.2% genetic diversity (Dingle et al. 2001), so even the 12.6% polymorphism remaining in ST-828 offers vastly superior resolution than is available with MLS. The reason for so much more of the polymorphism being homoplasious in ST-828 compared to the two *C. jejuni* STs (50% compared to 87%) is unknown, and open to speculation.

Table 5.4.1: Summary of sequence data for the three STs.

	ST-21	**ST-45**	**ST-828**
Genome length:	1529913	1401300	1289142
Non-polymorphic sites	1487590	1356574	1219277
Polymorphic sites	42323	44726	69865
Biallelic sites:	41378	43730	67547
Without missing data	27668	28093	24928
Compatible with ML tree	21875	22499	8821
Total sites used in analysis	**1515258**	**1379073**	**1228098**

5.4.2 Fine-scale Phylogenetic Structure within Sequence Types

For all three sequence types, samples isolated from different host species are often more closely related than those isolated from the same host species (Figure 5.4.1). If the isolates were host restricted, we would expect to see distinct clusters of the same coloured branches together. However, this is clearly not the case, with branches of the same colour – representing the reconstructed reservoir population of that lineage – scattered throughout the tree in all STs. Further, one might expect that isolates from mammals would be more closely related than those from avian species due to a more closely related physiology. Again, this is not true, with bird isolates in ST-45 and pig

isolates in ST-828 being closely related to both chicken and cattle isolates. This would suggest that the ability to colonise a specific host has either evolved several times throughout the tree, most plausibly through horizontal gene transfer given that mutation is rare, or that the isolates bear the innate ability to infect all types of host in the sample.

Much of the ancestral history of all lineages is inferred to have occurred within the chicken population, shown by the dominance of yellow branches. The MRCA of all three sequence types is inferred to be a chicken strain, with posterior probabilities of 0.611, 0.504 and 0.387 for ST-21, ST-45 and ST-828 respectively. Considering the ancestral dominance of chicken throughout the tree, these values show generally low support for the MRCA and thus must be taken with caution.

5.4.3 Rates of Zoonotic Transmission in *Campylobacter*

Under the host restricted hypothesis, isolates sampled from the same source reservoir will cluster into a single clade within the phylogeny. Since there are only two states in ST-21, this represents a minimum of one host migration on a branch within the tree. In ST-45 and ST-828, two migration events must occur under the same hypothesis, to result in three distinct clades. In fact, the total number of migration events for all three STs is estimated to be much higher than these minimum values, with 419.037 (99.629, 1168.663), 389.369 (82.112, 1286.315) and 31.463 (10.526, 321.469) migration events for ST-21, ST-45 and ST-828 respectively (Table 5.4.2). Since the minimum number of changes is not included in any of the credible intervals, this is strong evidence towards rejecting

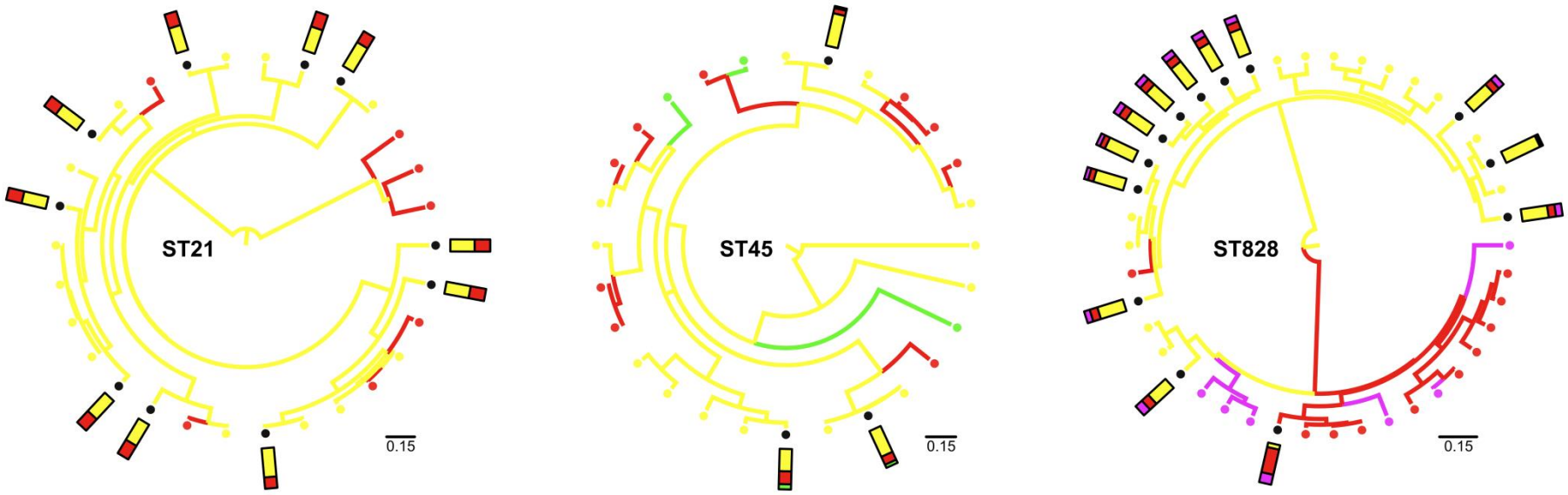


Figure 5.4.1: Ancestral source population for *Campylobacter*. Maximum clade credibility trees from the BEAST discrete trait analysis for a) ST-21, b) ST-45 and c) ST-828. Tips are coloured by host from which the sample was isolated: chicken (yellow), cattle (red), pig (pink), wild bird (green) and human (black). Branches are coloured according to the ancestral source population inferred using the maximum posterior probability. For each human case, the posterior probability of source is shown as a stacked barplot. Scale is given in units of coalescent time. A change in host may have occurred at any point on the branch, not necessarily at the node, and it is also possible to have a number of host switches occurring along a branch. Currently it is not possible to easily visualise either of these scenarios, and therefore this colouring represents the overall changes across the tree, rather than an exact representation of when and where switches occurred.

Table 5.4.2: Parameter estimates for each ST.

Parameter	ST-21	ST-45	ST-828
Substitution rate			
(10^{-3} /site /τ)	1.916	2.191	2.012
	(1.138, 3.112)	(1.263, 3.558)	(1.042, 3.200)
TMRCA			
Coalescent time, τ	0.880	0.868	0.649
	(0.384, 2.442)	(0.401, 2.587)	(0.228, 2.717)
Years	52.120	58.878	40.427
	(22.778, 144.855)	(27.201, 175.481)	(14.202, 169.245)
Host switching rate			
/τ	65.051	59.527	6.589
	(16.126, 98.373)	(13.486, 97.946)	(1.640, 60.836)
/year	1.097	0.878	0.106
	(0.272, 1.658)	(0.199, 1.444)	(0.026, 1.009)
Number of migrations			
	419.037	389.369	31.463
	(99.629, 1168.663)	(82.112, 1286.315)	(10.526, 321.469)
Estimated host frequencies			
Chicken	0.614	0.506	0.356
	(0.384, 0.812)	(0.316, 0.694)	(0.071, 0.681)
Cattle	0.386	0.341	0.324
	(0.188, 0.616)	(0.176, 0.535)	(0.104, 0.619)
Wild bird	-	0.140	-
		(0.043, 0.312)	
Pig	-	-	0.290
			(0.102, 0.602)
Relative migration rates between host species			
Chicken-Cattle	0.692	0.770	0.346
	(0.026, 3.699)	(0.034, 3.748)	(0.010, 2.359)
Chicken-Wild Bird	-		-
		0.612	
Chicken-Pig	-	(0.022, 3.333)	0.335
		-	(0.010, 2.443)
Cattle-Wild Bird	-		-
		0.751	
Cattle-Pig	-	(0.042, 3.674)	1.561
		-	(0.232, 5.164)

Using the mutation rate 3.23×10^{-5} substitutions per site per year (Wilson et al. 2009) for calibration with the median evolutionary rate from the BEAST analysis, one unit of coalescent time (τ) is equal to: 59.318 years in ST-21, 67.832 years in ST-45, and 62.291 years in ST-828. The median value was used for estimating all other parameters in years.

the host specific hypothesis in favour of the STs being generalists. However, it is noted that the number of migration events in ST-21 and ST-45 is over ten times that of ST-828, suggesting that there may be some difference between the *C. jejuni* and *C. coli* strains. This pattern is also seen in the overall migration rate, where again the rate is in the order of nine to ten times higher in ST-21 and ST-45 than in ST-828. Using a mutation rate of 3.23×10^{-5} substitutions per site per year (Wilson et al. 2009) for calibration, this corresponds to approximately one host migration event every nine years for ST-828 compared to one every 12 or 13 months in ST-21 and ST-45.

However, the credible intervals are wide, particularly for the *C. jejuni* sequence types, showing that there is much uncertainty in the estimates. It is important to note that the posterior distribution for the overall migration rate for ST-21 and ST-45 does not have a peak value, and the highest posterior density (HPD) interval contains the maximum value. Other priors with a wider distribution were tried, but this observation continued to be the case. This suggests that the true value tends to infinity, and that the rate reported here is an underestimate. Thus, I propose that the rate of migration is so rapid in ST-21 and ST-45 that there is no association between genetic structure and host species, and thus that isolates are equally able to transmit between species as within them. Statistically, this is equivalent to randomly assigning individuals to host species according to their probability in the original sample, which is illustrated by the estimated host frequencies being very close to the frequency of hosts in the original sample (Table 5.4.2). This adds to the evidence of these sequence types being true generalists.

5.4.4 Tracing the Source of Human Infection

In Figure 5.4.1, the human cases at the tips of the tree are represented with black circles, and the posterior probability of the sources for the human isolates are illustrated by the bar plots and also given in Table 5.4.3. It is clear that in ST-21, ST-45 and ST-828 all but one of the 26 human cases were attributed to a chicken source (96%), and the remaining human case was found to be most likely from a cattle source (PP: 63.7%). Whilst this fits with results found using MLST data (for example the findings of Wilson et al. (2008), Sheppard et al. (2009) and Mullner et al. (2009)), there remains much uncertainty in the estimates. Even when human isolates are most closely related to chicken-only clades, there is 30-40% uncertainty in the host allocation. This is most likely to be due to the high switching rates noted in the previous section, and likely also alludes to the predominance of chicken isolates in the original sample. However, this uncertainty is still comparable to the 36% error rate of correctly assigning isolates to their source population using the model of Wilson et al. (2008) with MLST data, and in some human cases gives much more precise estimates.

Table 5.4.3: Posterior probabilities for the source of human infection.

		Posterior Probability of Source			
Isolate Number	**ST**	**Chicken**	**Cattle**	**Bird**	**Pig**
29	21	60.91	39.09	-	-
117	21	61.13	38.87	-	-
116	21	71.95	28.05	-	-
182	21	60.97	39.03	-	-
189	21	61.71	38.29	-	-
195	21	61.00	39.00	-	-
36	21	64.86	35.14	-	-
34	21	64.11	35.89	-	-
37	21	61.17	38.83	-	-
60	21	68.23	31.77	-	-
55	45	81.73	13.04	5.23	-
119	45	70.05	21.48	8.47	-
32	45	55.19	31.92	12.89	-
159	828	11.24	63.70	-	25.06
170	828	57.80	21.34	-	20.85
169	828	65.85	19.78	-	14.37
161	828	76.49	13.68	-	9.83
162	828	76.52	13.62	-	9.86
160	828	64.71	19.88	-	15.41
163	828	64.74	19.86	-	15.40
167	828	64.69	19.92	-	15.39
156	828	64.64	19.98	-	15.38
168	828	64.70	19.92	-	15.38
165	828	66.58	17.87	-	15.56
2	828	91.98	4.10	-	3.93
19	828	65.32	18.38	-	16.31

5.5 Discussion

This chapter investigated the source of *Campylobacter* infection in humans using whole genome sequencing, focusing on STs that appeared to be weakly host-associated on the basis of MLST. I tested whether these STs were in reality aggregates of strong host-

restricted sub-groups, or whether they represented genuine generalists (Gripp et al. 2011). I estimated rates of zoonotic transfer between *Campylobacter* reservoir populations, and attributed individual human cases to potential animal sources.

In summary, I found fine-scale genetic structure below the level of ST that does not, however, appear to be host associated. I therefore concluded that these strains represent genuine generalists, adapted to live on multiple host species, and freely transmitting zoonotically between animal hosts. This flexible lifestyle may represent a niche in its own right in environments such as farms, where multiple mammalian and avian species routinely come into close contact.

The picture is slightly different for the *C. coli* ST-828 strain, where there was evidence for relatively slower rates of zoonosis, compared to the *C. jejuni* ST-21 and ST-45 strains (Table 5.4.2). Even in ST-828 however, I was able to strongly reject the hypothesis of a unique host jump founding the population in each new species, evidenced by the scattering of isolates sampled from different sources throughout the phylogeny.

With the rapid host-switching even below the level of individual STs, and the sparse sampling of potential source isolates compared to MLST studies (in which hundreds to thousands of isolates were available (Wilson et al. 2008; Mullner et al. 2009; Sheppard et al. 2009)), there was no clear additional information regarding source of infection provided by the whole genome sequence in the majority of human cases. This tells us that there is a fundamental trade-off between the number of samples and the number of

loci sequenced, and that the approximately thousand-fold increase in sequence information afforded by whole genomes over MLST does not on its own trump the need for detailed sampling of the potential source populations. The one exception to this trend was in ST-828, where I was able to confidently attribute a human case to a cattle, rather than a chicken, source (Table 5.4.3). It is fair to say that while this case provides evidence of the potential for whole genome sequencing to provide greater resolution in tracking the source of infection in individual cases of *Campylobacter* infection, there is no substitute for intensive sampling of source populations.

My results show that cross-species transmission occurred at different rates in different STs, with far more frequent zoonotic transfer in ST-21 and ST-45 than in ST-828. I estimated that there were 419 migration events across the tree in ST-21 and 389 in ST-45, compared to only 31 in ST-828, in all cases significantly greater than the minimum number of migration events required under hypothesis of strong host restriction (Table 5.4.2). In addition, the overall migration rate was around once per year in ST-21and ST-45, with a strong possibility that this was an underestimate, but only 0.1 in ST-828. Even taking into account the effect of total tree length on the number of migration events, there was clearly a different pattern of host migration between the *C. jejuni* and *C. coli.* sequence types.

The cause of the different overall rates of migration can be explained by the relative rates of transitions between species. In ST-45, there was very little difference in the relative rates of transmission between hosts (Table 5.4.2). This similarity in migration

rates between different host species suggests that it is irrelevant what type of host transmission event is taking place. However, in ST-828, switches are nearly five times more frequent between cattle and swine compared with the rates of exchange between these species and chicken. This distinction between inter-mammalian and mammalian-avian exchange rates suggests that zoonosis is fundamentally different in ST-828 compared to ST-21 or ST-45. This may be due to less a restrictive route of transmission between the mammalian species, either through opportunity, (say, the two mammal species may be more likely to share the same environment that ST-828 isolates occupy), or through affinity (ST-828 isolates may be more specialized to mammalian versus avian hosts than ST-21 and ST-45 isolates). The shorter genome of ST-828 compared to ST-21 and ST-45 suggests that the latter is a distinct possibility.

Recently Sheppard et al. (2013b) used a genome wide association study to show that the genes required for vitamin B_5 biosynthesis show an association with cattle, and are more variable with regards to presence in isolates from chicken. They suggest that this may be due to the levels of vitamin B_5 being lower in grass, the main diet of cattle, than in the cereals and grains that constitute the main chicken diet. This provides circumstantial support for a possible genomic basis for the difference in host specialism versus generalism in ST-828 versus ST-21 and ST-45.

The effect of recombination even within individual *Campylobacter* STs is profound, and is challenging for phylogenetic inference. Several steps were taken to overcome the effects of recombination, including the removal of homoplasious sites incompatible with

the maximum likelihood tree to leave only sites representing the dominant underlying phylogeny. This led to the removal of up to 82% of the polymorphism in the case of ST-828 (Table 5.4.1). Discarding data is obviously not ideal, however it is important to note that the remaining resolution is still far greater than MLST, and overcomes the issue of recombination which otherwise would be highly problematic in a BEAST analysis. There are software available (for example ClonalFrame (Didelot and Falush 2007)) that allow ancestral relationships to be inferred alongside recombination events. However, such methods tend to be limited in their model choice (for example, regarding a changing population size), and are often computationally slow, especially for large numbers of genomes. Unlike ClonalFrame, the approach described in this chapter identifies all homoplasious sites, with the distinct advantage that it allows for a range of phylogenetic analyses to be performed.

The issues surrounding recombination have plagued phylogenetic inference in bacterial genomics in the whole genome sequencing era, so this general approach could pave the way for many new applications. In addition to investigating zoonotic transmission in numerous other species, the migration model could be used to compare categories of human hosts – for example, carriage and disease-causing isolates, or differences in hospital and community samples. In addition to migration, the dissipation of phenotypes could be compared throughout a phylogeny, including antibiotic resistance and symptoms of disease.

All but one of the human isolates were allocated to a chicken source; the remaining isolate, from ST-828, was attributed to a cattle source (Table 5.4.3). This supports the conclusion of numerous previous studies, including both those using MLST to investigate the source population, and epidemiological studies, that campylobacters attributable to chicken account for the majority of human illness (Harris et al. 1986; Wingstrand et al. 2006; Wilson et al. 2008; Mullner et al. 2009; Sheppard et al. 2009). This reiterates the importance of measures aimed at controlling food-borne disease in the agricultural industry. At the same time, this result suggests that without equally intensive sampling of *Campylobacter* reservoirs as undertaken in previous MLST studies, whole genome sequencing does not add sufficiently greater resolution to unambiguously attribute the source of individual human cases.

An important limitation of the approach is that it is only possible to attribute human cases to sources already present in the sample – however distantly related they might actually be. Both the range of species and number of isolates collected from those species were modest in this analysis, although the former was at least informed by previous MLST work. Since, in phylogenetic terms, the inferred source population of a human isolate will be that of its closest non-human relative, the tips need to be very densely sampled to get the most accurate results. The high zoonosis rates estimated here underline this point. Ideally, the make-up of a sample needs to reflect that of the bacterial population at large, but this aspiration ignores practical concerns such as the availability of access to sampling locations, personnel and sequencing budget. As these

challenges are overcome, this phylogenetic method will shed more light on the transmission dynamics of *Campylobacter* since it is easily scalable for larger datasets.

Chapter 6: Summary

Infectious disease represents a burden on modern society, impacting both individual health and the economy. Whole genome sequencing has the power to revolutionise how we understand the dynamics of infectious disease and its management (Chan et al. 2012; Wilson 2012; Didelot 2013). Patterns of diversity between isolates can be used to reconstruct evolutionary relationships within and between outbreaks, allowing us to track the epidemiology from a different perspective. In this thesis, I have presented approaches that exploit the greater resolution afforded by genomic sequencing to answer important questions about the evolution and transmission of three major human pathogens: hepatitis C, norovirus and *Campylobacter*.

In this final chapter, I first summarise the key findings for each results chapter, before highlighting the overarching themes and limitations of the thesis. I then take a step back to put this work into the broader context, and survey the impact of future developments in the field.

6.1 Thesis Summary

6.1.1 Coalescent Inference for Infectious Disease

In Chapter 3, I investigated whether differences in genetic diversity between hepatitis C virus (HCV) epidemics could be explained by the underlying epidemiological processes. I found that the age of the epidemic had the strongest predictive power, followed by population density and subtype (Table 3.5.2). I dated all but one of the epidemics to

within the last century (Table 3.5.1), reiterating the importance of 20th century phenomena such as blood transfusion and needle sharing in the spread of HCV (Shepard et al. 2005). Model selection showed that pathogen effective population sizes have increased exponentially in keeping with the SI model in most epidemics, but in three localised epidemics more complex SIR dynamics were detected (Table 3.5.3). For these epidemics, I was able to estimate the prevalence of infection, basic reproductive number and mean duration of infection from genetic data alone (Table 3.5.4). Despite the chronic nature of HCV, my results suggest that most secondary transmission occurs shortly after infection, with a much shorter duration of infection (1.24-1.55 years) than previously estimated (Pybus et al. 2001; Lavanchy 2011).

The analysis in Chapter 3 was based on combined epidemiological-coalescent models implemented in BEAST. This approach allows epidemiological parameters such as the intrinsic growth rate to be estimated using genetic data alone. This has advantages over traditional methods when comprehensive monitoring during an outbreak is infeasible but contemporary sampling of isolates for sequencing is possible, allowing a direct comparison of growth rates between outbreaks to identify common patterns, and investigate the effect of any interventions taking place (for example, the introduction of a vaccine or treatment). The approach is also readily extendable, allowing greater biological complexity to be incorporated in future analyses as and when required. The meta-population model under a more general formulation permits heterogeneity in host types (Wakeley and Aliacar 2001), enabling the investigation, for example, of super shedders within outbreak situations. These mean that the model could easily be applied

to a range of pathogens, allowing genetic analyses to complement epidemiological approaches in the prediction and comparison of real-time outbreaks.

I introduced a model averaging approach in Section 3.4.2, to help overcome the problems associated with using the harmonic mean estimator to estimate the marginal likelihood for the Bayes factor test (Xie et al. 2011). Two other marginal likelihood estimators have since been implemented in BEAST: stepping-stone sampling and path sampling (Baele et al. 2012). Like the model averaging approach I presented here, these can be used to compare models of molecular evolution and demographic change. However, in order to calculate the marginal likelihood, these methods require a second MCMC based upon the posterior from the BEAST run and thus remain computationally intense. A comparison of all three methods was out of the scope of this thesis, though it would be interesting to investigate how they compare in terms of both accuracy and total computational time.

6.1.2 Evolution and Transmission of GII.4 Norovirus

In Chapter 4, my analysis focused on the evolution and transmission of GII.4 norovirus. This single genotype is responsible for more than 80% of global epidemics, and is an enormous burden to health services (Cooke et al. 2003; Lopman et al. 2005; Bull and White 2011). Emerging strains were found to be most closely related to strains circulating two or more seasons previously, rather than the previously circulating strain Figure 4.4.1b. I estimated the genome-wide rate of evolution at approximately 1 substitution every 9 days, which is in line with estimates calculated from either the

capsid or the partial RdRp (Bok et al. 2009; Bull et al. 2010; Siebenga et al. 2010). Using a stochastic model of transmission, I found that up to 86% of infection in Oxford University Hospitals from 2009-2013 was attributable to another patient in the same ward, mainly due to transmission attributable to another patient in the hospital (Table 4.4.3). The rate at which patients entered the hospital infected with norovirus was very low, estimated at less than 0.05 infections per 10,000 patient days (Table 4.4.2). In contrast, sharing a ward with an infectious patient increased the transmission risk 100-fold above background hospital risk, highlighting the importance of ward management during outbreaks. Knowledge of this sort can prove valuable in assessing the efficacy of outbreak investigation and control measures.

Direct transmission in the nosocomial spread of norovirus has long been thought important (Teunis et al. 2013). However, the stochastic model applied to GII.4 norovirus provided unprecedented depth, inferring the relative risk of alternate transmission routes and allowing prediction of the impact of future outbreaks. The stochastic transmission model used here could be readily extended to find the relative transmission risk between ward types, which could provide insights in to the way norovirus is controlled – for example informing different strategies in according to wards that are at most risk. The model is not just limited to norovirus; it could be applied to other nosocomial diseases, including *Staphylococcus aureus*, *Escherichia coli* and enterococci, to learn about disease incidence and dominant transmission routes.

As part of the analysis in Chapter 4, I estimated the most up-to-date genome-wide evolutionary clock rate yet reported (Section 4.4.1). Previous studies have focused on the VP1 region alone, due to its role in host cell binding and viral antigenicity, but this represents only represents 21% of the genome (Bok et al. 2009; Bull et al. 2010; Siebenga et al. 2010). In this chapter, I summarised the evolutionary rate over the entire genome, despite the capsid area being thought of as more diverse than the rest of the genome. Future work could investigate variation in the clock rate across the genome by using hierarchal phylogenetic models. These are now available in BEAST, and can be used to analyse sequences that have been partitioned (by ORF, for example). Hierarchal phylogenetic models allow segments of the genome to have different topologies and allow better precision of estimates within partitions by pooling information, but also allow for overall estimates across partitions (i.e. rates for individual ORFs in addition to an average across the whole genome) (Suchard et al. 2003).

6.1.3 Zoonotic Transmission of *Campylobacter*

In the final results chapter, I investigated zoonosis and the source of human disease with *Campylobacter.* The increased resolution of whole genome sequencing revealed fine-scale genetic structure below the level of ST, but this did not appear to be host associated (Figure 5.4.1). I concluded that ST-21, ST-45 and ST-828 represent genuine generalists, adapted to live in multiple animal hosts, and able to freely transmit zoonotically between them. I detected variation in cross-species transmission rates, with far more frequent zoonotic transfer in *C. jejuni* strains ST-21 and ST-45 than in the *C. coli* strain ST-828 (Table 5.4.2). In ST-828, the rate of zoonosis between cattle and swine was

nearly five times that of cattle and chicken or swine and chicken, whereas in ST-45, there was very little difference in the relative rates between avian and mammal species. Although 24 of the 25 human cases were attributed to chicken (Table 5.4.3), the ability to detect a specific case of cattle-associated human infection in ST-828 demonstrates the potential impact and extraordinary resolution of whole genome sequencing on source attribution in *Campylobacter* and elsewhere.

The number of sequences used in this *Campylobacter* study (30, 28 and 42 of ST-21, ST-45 and ST-828 respectively) was small compared to the thousands of sequences that have been used in previous studies using MLST (Wilson et al. 2008; Mullner et al. 2009; Sheppard et al. 2009). Nevertheless, the method I developed for reconstructing zoonotic transmission is readily extendable to much larger datasets, and this would be the logical next step in examining zoonosis in the three *Campylobacter* STs studied here. First, whole genome sequencing of more intensively sampled isolates is required, covering all possible reservoirs of disease, in order to obtain the most accurate rates of zoonosis and attribution for the source of human disease.

The MLST collections for *Campylobacter* (www.pubmlst.org/campylobacter, Jolley and Maiden (2010)) are a rich resource, not only for MLST of each isolate, but also information including the host species, type of disease presented (e.g. carrier, gastroenteritis), isolation date, and known epidemiology (whether the isolate was a sporadic case, or part of a large outbreak). Assuming whole genomes follow this lead as more of them become available, the analysis presented in Chapter 5 could easily be

extended to take into account this added information, for example, using the time of sampling, to better calibrate evolutionary time with days, months and years. If sampling is deep enough, there is the potential to narrow down when (and even where) migration events most likely took place throughout the tree, making it possible to elucidate whether it is genetic or environment barriers having an effect on mammal-avian transmission in ST-828.

6.2 Uniting Themes

The methods that I have developed and applied in this thesis represent tentative steps towards fully exploiting the potential of ever-increasing volumes of epidemiological and pathogen genome data to make inference about infectious disease dynamics. The field is rapidly growing, and there is a trend towards work, such as that presented in this thesis, that jointly considers evolution and transmission side by side (Hedge et al. 2013; Jombart et al. 2014; Kühnert et al. 2014; Rasmussen et al. 2014). Each of the methods considered in this thesis was developed with the aim of contributing new tools for understanding infectious disease. However, the assumptions that were necessary to make progress in each setting has, to a greater or lesser extent, rendered the methods somewhat specific to particular problems – the deterministic metapopulation models in Chapter 3 are appropriate for understanding population-level dynamics such as the emergence of new pathogens over decades, but do not account for inherent stochasticity at the level of person-to-person transmission, an important force when considering norovirus transmission in hospitals, for example (Chapter 4). Currently, there is no one 'catch-all' approach as there are in other areas, for example in phylogenetics where

flexible software such as BEAST caters for a spectrum of inferential problems united within a coalescent-based framework. The lack of a general toolbox currently limits the accessibility of joint epidemiological and evolutionary inference for specialists less familiar with genetic epidemiological inference, but for whom these sorts of model would hold useful information, such as public health epidemiologists and clinicians. A solution to this problem may be some way off, but the challenge represents an exciting opportunity to strive towards a goal that, if met, would likely have far-reaching benefits.

A recurring difficulty in this thesis, and indeed the field at large, is the question of how to account for the effect of recombination on phylogenetic analyses. Homologous recombination is a major driver of pathogen evolution (Didelot et al. 2012b) that has several consequences for phylogenetic reconstruction. The differing ancestral histories of segments in the genome cannot be represented on a single genealogy, and instead, require an Ancestral Recombination Graph (ARG) or phylogenetic network (Hudson and Kaplan 1988; Griffiths and Marjoram 1996; Huson and Bryant 2006). A common approach is to brush aside the question of recombination, and apply phylogenetic methods to pathogen genomes irrespective of detectable levels of recombination. This is a clear instance of model mis-specification, and its effects have been investigated, with Schierup and Hein (2000) showing that recombination skews branch lengths and growth rates, making phylogenies appear star-like, and also leads to an overestimate in the substitution rate heterogeneity and loss of molecular clock. Recombination was a clear issue in the analyses both of hepatitis C virus and *Campylobacter*. In these chapters, I followed a common approach which is to remove data (sequences or regions of the

genome) with evidence of recombination. In Chapter 3, this led to the removal of six collections of sequences from the meta-analysis (Table 3.4.1), and in Chapter 5, the removal of homoplasious sites left only 12.6% of the original polymorphism in ST-828 (Table 5.4.1). Although in both cases there was still plenty of data available from which to draw conclusions, it would be preferable to use all the data available. In Chapter 4, the stochastic model was defined by sequence types, and thus did not require recombinant samples to be removed. However, the presence of recombinants in the 2012-13 season (Table 4.4.1) did highlight that two diverse isolates need to come into contact for recombination to have taken place, which plausibly represents the convergence of two transmission chains in a single patient.

As genome sequencing becomes more widespread, it is likely that a history of recombination will be detected in more and more pathogens. Indeed, even over the course of my time as a D.Phil. student, the importance of recombination in norovirus has begun to be revealed by influential papers showing how recombination in the ORF1/ORF2 may facilitate emergence of new strains (Mathijs et al. 2011; Eden et al. 2013; Fonager et al. 2013; Martella et al. 2013; Wong et al. 2013). This is a direct result of there being more whole genome sequences publically available. Current methods for inferring the ancestral history of a sample, such as ClonalFrame (Didelot and Falush 2007) and SMARTIE (Bloomquist and Suchard 2010) need to become faster and offer a wider scope of models in response, with room for the development of methods that can cope with even more data.

6.3 Future Directions

There have been dramatic advances in phylogenetic methodology and sequence availability which have taken place during the course of my D.Phil studies, and the pace of change shows no sign of slowing. In the following, I consider recent developments to the field as whole, and how my work corresponds to these, with views to the future.

Perhaps the most striking trend is the continuing decreasing costs and turnaround times for whole genome sequences (Nature 2010). Much of the focus has been on sequencing the human genome, but of greater relevance to this thesis is the potential for whole genome sequencing to transform clinical microbiology (Pallen et al. 2010; Didelot et al. 2012a; Loman et al. 2012; Wilson 2012; Didelot 2013). It is now possible to track the transmission of infectious disease within patients (Cramer et al. 2011; Young et al. 2012), at a ward, hospital and community level (Rohde et al. 2011; Eyre et al. 2012; Köser et al. 2012), and relate these outbreaks to global dynamics. Whilst such studies have tended to be localised within research institutions in the past, new nationwide initiatives such as the UK 100K genomes project (http://www.genomicsengland.co.uk/) will widen the scope and impact of sequencing on clinical microbiology. This extends the concept of 'personalised medicine' to being personal to an individual's microbiome as well as their human genome.

High throughput sequencing has already begun to change microbiology, but new technologies are pushing the revolution further. In 2012, Oxford Nanopore Technologies announced a new sequencing platform in which a molecule of replicating DNA is passed through a protein nanopore, and the nucleotides are read using changes in electrical conductivity (Check Hayden 2012). It is suggested that 200-400 bases can be read per second, with contiguous read lengths of up to tens of kilobases (Didelot et al. 2012a). Whilst this has not yet been released commercially (although an early access program has just been rolled out), reads lengths ten to one thousand times longer than currently achieved by industry leader Illumina have the potential to completely overcome the difficulties associated with genome assembly, which struggles to produce completely closed genomes in the presence of repetitive regions (Zerbino and Birney 2008; Iqbal et al. 2012). The ability to quickly obtain high-quality closed genomes would speed up the computational pipeline required to go from sampling to the sorts of analyses presented in this thesis (Loman et al. 2012), and make accessible repetitive parts of the genome that are currently out of reach of short-read sequencing platforms. This is particularly important because many genes of paramount importance in host-pathogen interactions, such as adhesins and toxins, are typically repetitive (Giesemann et al. 2008; Linhartová et al. 2010). Among other potential benefits is the prospect of untangling mixed samples more easily, paving the way for an improved understanding of within-host dynamics.

Most striking is the dawning reality of taking a sample from a patient, and having genomic results available a few hours later – a stark contrast to current practice that

often requires the culture of slow-growing pathogens (Didelot et al. 2012a; Köser et al. 2012). These results can shed light on the species identity of the aetiological agent, possible barriers to treatment (for example, notification of any resistance mutations), early warning of virulent strains, and the relationship to other cases in hospital (for example, whether cases on same ward are the result of direct transmission). However, these results need to be readily accessible to non-specialists if they are to be used on a routine basis at the local, national and global level. It is hoped that the methodologies developed in this thesis represent a step towards this goal by facilitating joint epidemiological and genomic inference, and that they will be used as part of a wider program of understanding how joint genetics and epidemiological modelling can further our knowledge of disease dynamics.

6.4 Final Remarks

As genomic sequencing of patient samples becomes routine in hospital microbiology and public health laboratories, the ability to integrate population genetic inference and epidemiology will become only more valuable in the fight against infectious disease. The work presented in this thesis shows how the vast amounts of data generated as a result can be exploited to understand the transmission of disease, at the global, local and individual levels. Ultimately, it is hoped that this knowledge will be used to improve outbreak investigation, enhance individual patient management, and inform public health strategies.

Literature Cited

Adams N (2013) Norovirus: early start to norovirus season 2012/13. GEZI Quarterly 1:3–4.

Adessi C, Matton G, Ayala G, et al. (2000) Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. Nucleic Acids Res 28:E87.

Adler JL, Zickl MD (1969) Winter vomiting disease. J Infect Dis 119:668–673.

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Contr 19:716–723.

Akkarathamrongsin S, Praianantathavorn K, Hacharoen N, et al. (2010) Geographic Distribution of Hepatitis C Virus Genotype 6 Subtypes in Thailand. J Med Virol 262:257–262. doi: 10.1002/jmv.21680

Allain J-P (2003) Transfusion risks of yesterday and of today. Transf Clin Biol 10:1–5. doi: 10.1016/S1246-7820(02)00276-8

Allos BM (2001) *Campylobacter jejuni* infections: update on emerging issues and trends. Clin Infect Dis 32:1201–6. doi: 10.1086/319760

Alter HJ, Purcell RH, Holland P V, et al. (1981) Donor transaminase and recipient hepatitis. J Am Med Assoc 952:630–4.

Alter HJ, Seeff LB (2000) Recovery, persistence, and sequelae in hepatitis C virus infection: a perspective on long-term outcome. Semin Liver Dis 20:17–35.

Anderson RM, May RM (1991) Infectious diseases of humans: dynamics and control. Oxford Science Publications, Oxford

Anderson S (1981) Shotgun DNA sequencing using cloned DNase I-generated fragments. Nucleic Acids Res 9:3015–3027.

Ando T, Jin Q, Gentsch JR, et al. (1995) Epidemiologic applications of novel molecular methods to detect and differentiate small round structured viruses (Norwalk-like viruses). J Med Virol 47:145–52.

Aoki Y, Suto A, Mizuta K, et al. (2010) Duration of norovirus excretion and the longitudinal course of viral load in norovirus-infected elderly patients. J Hosp Infect 75:42–6. doi: 10.1016/j.jhin.2009.12.016

Arenas M, Valiente G, Posada D (2008) Characterization of reticulate networks based on the coalescent with recombination. Mol Biol Evol 25:2517–20. doi: 10.1093/molbev/msn219

Aris-Brosou S, Yang Z (2002) Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. Syst Biol 51:703–14. doi: 10.1080/10635150290102375

Aslam M, Aslam J, Mitchell BD, Munir KM (2005) Association between smallpox vaccination and hepatitis C antibody positive serology in Pakistani volunteers. J Clin Gastroenterol 39:243–6.

Atmar RL, Bernstein DI, Harro CD, et al. (2011) Norovirus vaccine against experimental human Norwalk Virus illness. N Engl J Med 365:2178–87. doi: 10.1056/NEJMoa1101245

Atreya CD (2004) Major foodborne illness causing viruses and current status of vaccines against the diseases. Foodb Pathog Dis 1:89–96. doi: 10.1089/153531404323143602

Avise JC, Shapira JF, Daniel SW, et al. (1983) Mitochondrial DNA differentiation during the speciation process in Peromyscus. Mol Biol Evol 1:38–56.

Bacon BR, Gordon SC, Lawitz E, et al. (2011) Boceprevir for previously treated chronic hepatitis C virus genotype 1 infection. N Engl J Med 364:1207–1217. doi: 10.1056/NEJMoa100948

Baele G, Lemey P, Bedford T, et al. (2012) Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol Biol Evol 29:2157–67. doi: 10.1093/molbev/mss084

Bahl J, Nelson MI, Chan KH, et al. (2011) Temporally structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans. Proc Natl Acad Sci USA 108:19359–64. doi: 10.1073/pnas.1109314108

Ballesteros ÁL, Franco S, Fuster D, et al. (2004) Early HCV dynamics on Peg-interferon and ribavirin in HIV/HCV co-infection: indications for the investigation of new treatment approaches. AIDS 18:59–66. doi: 10.1097/01.aids.0000104370.21567.a3

Baltimore D (1971) Expression of animal virus genomes. Bacteriol Rev 35:235–41.

Bari A, Akhtar S, Rahbar MH, Luby SP (2001) Risk factors for hepatitis C virus infection in male adults in Rawalpindi-Islamabad, Pakistan. Trop Med Int Heal 6:732–8. doi: 10.1046/j.1365-3156.2001.00779.x

Barker J, Jones M V (2005) The potential spread of infection caused by aerosol contamination of surfaces after flushing a domestic toilet. J Appl Microbiol 99:339–47. doi: 10.1111/j.1365-2672.2005.02610.x

Barker J, Vipond IB, Bloomfield SF (2004) Effects of cleaning and disinfection in reducing the spread of norovirus contamination via environmental surfaces. J Hosp Infect 58:42–49. doi: 10.1016/j.jhin.2004.04.021

Barnard M, Albert H, Coetzee G, et al. (2008) Rapid molecular screening for multidrug-resistant tuberculosis in a high-volume public health laboratory in South Africa. Am J Respir Crit Care Med 177:787–92. doi: 10.1164/rccm.200709-1436OC

Barnes I, Matheus P, Shapiro B, et al. (2002) Dynamics of Pleistocene population extinctions in Beringian brown bears. Science 295:2267–70. doi: 10.1126/science.1067814

Barrell BG, Sanger F (1969) The sequence of phenylalanine tRNA from *E. coli*. FEBS Lett 3:275–278.

Bassett SE, Brasky KM, Lanford RE (1998) Analysis of hepatitis C virus-inoculated chimpanzees reveals unexpected clinical profiles. J Virol 72:2589–99.

Bateson W (1894) Materials for the study of variation treated with especial regard to discontinuity in the origin of species. Macmillan and Co., New York

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

Bayes M, Price M (1763) An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. Philos Trans R Soc London 53:370–418. doi: 10.1098/rstl.1763.0053

Bedford T, Cobey S, Pascual M (2011) Strength and tempo of selection revealed in viral gene genealogies. BMC Evol Biol 11:220. doi: 10.1186/1471-2148-11-220

Behrens SE, Tomei L, De Francesco R (1996) Identification and properties of the RNA-dependent RNA polymerase of hepatitis C virus. EMBO J 15:12–22.

Beijerinck MW (1898) Ueber ein Contagium vivurn fluidum als Ursaehe der Fleekenkrankheit der Tabaksblätter. Verh K Akad Wet Amsterdam 65:3–21.

Benedictow OJ (2004) The Black Death 1346-1353: The Complete History. The Boydell Press, Woodbridge

Benzer S, Champe SP (1961) Ambivalent rII mutants of phage T4. Proc Natl Acad Sci USA 47:1025–38.

Berg HC, Anderson RA (1973) Bacteria swim by rotating their flagellar filaments. Nature 245:380–2. doi: 10.1038/245380a0

Biggs PJ, Fearnhead P, Hotter G, et al. (2011) Whole-genome comparison of two *Campylobacter jejuni* isolates of the same sequence type reveals multiple loci of different ancestral lineage. PLoS One 6:e27121. doi: 10.1371/journal.pone.0027121

Bissig K-D, Wieland SF, Tran P, et al. (2010) Human liver chimeric mice provide a model for hepatitis B and C virus infection and treatment. J Clin Invest 120:924–930. doi: 10.1172/JCI40094D

Black RE, Levine MM, Clements ML, et al. (1988) Experimental *Campylobacter jejuni* infection in humans. J Infect Dis 157:472–9.

Blackshields G, Wallace IM, Larkin M, Higgins DG (2006) Analysis and comparison of benchmarks for multiple sequence alignment. In Silico Biol 6:321–39.

Blancou J, Chomel BB, Belotto A, Meslin FX (2005) Emerging or re-emerging bacterial zoonoses: factors of emergence , surveillance and control. Vet Res 36:507–522. doi: 10.1051/vetres:2005008

Blaser MJ (1997) Epidemiologic and clinical features of *Campylobacter jejuni* infections. J Infect Dis 176:S103–5. doi: 10.1086/513780

Bloom BR, Murray CJ (1992) Tuberculosis: commentary on a reemergent killer. Science 257:1055–64.

Bloomquist EW, Suchard MA (2010) Unifying vertical and nonvertical evolution: a stochastic ARG-based framework. Syst Biol 59:27–41. doi: 10.1093/sysbio/syp076

Bode LGM, van Wunnik P, Vaessen N, et al. (2012) Rapid detection of methicillin-resistant *Staphylococcus aureus* in screening samples by relative quantification between the *mecA* gene and the SA442 gene. J Microbiol Methods 89:129–32. doi: 10.1016/j.mimet.2012.02.014

Bogen SL, Pan W, Ruan S, et al. (2009) Toward the back-up of boceprevir (SCH 503034): discovery of new extended P_4-capped ketoamide inhibitors of hepatitis C virus NS3 serine protease with improved potency and pharmacokinetic profiles. J Med Chem 52:3679–88. doi: 10.1021/jm801632a

Bok K, Abente EJ, Realpe-Quintero M, et al. (2009) Evolutionary dynamics of GII.4 noroviruses over a 34-year period. J Virol 83:11890–901. doi: 10.1128/JVI.00864-09

Bos KI, Stevens P, Nieselt K, et al. (2012) *Yersinia pestis*: new evidence for an old infection. PLoS One 7:e49803. doi: 10.1371/journal.pone.0049803

Boyer M, Madoui M-A, Gimenez G, et al. (2010) Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4th domain of life including giant viruses. PLoS One 5:e15530. doi: 10.1371/journal.pone.0015530

Bromham L, Penny D (2003) The modern molecular clock. Nat Rev Genet 4:216–24. doi: 10.1038/nrg1020

Brownlee GG, Sanger F, Barrell BG (1967) Nucleotide sequence of 5S-ribosomal RNA from *Escherichia coli*. Nature 215:735–736. doi: 10.1038/215735a0

Bryant D (2003) A classification of consensus methods for phylogenetics. DIMACS Series in Discrete Mathematics and Theoretical Computer Science 61:163–184.

Buchanan RE, St. John-Brooks R, Breed RS (1947) International bacteriological code of nomenclature. J Bacteriol 55:287–306.

Bull RA, Eden J-S, Rawlinson WD, White PA (2010) Rapid evolution of pandemic noroviruses of the GII.4 lineage. PLoS Pathog 6:e1000831. doi: 10.1371/journal.ppat.1000831

Bull RA, Hansman GS, Clancy LE, et al. (2005) Norovirus recombination in ORF1/ORF2 overlap. Emerg Infect Dis 11:1079–1085.

Bull RA, Tu ET V, McIver CJ, et al. (2006a) Emergence of a new norovirus genotype II.4 variant associated with global outbreaks of gastroenteritis. J Clin Microbiol 44:327–333. doi: 10.1128/JCM.44.2.327

Bull RA, White PA (2011) Mechanisms of GII.4 norovirus evolution. Trends Microbiol 19:233–40. doi: 10.1016/j.tim.2011.01.002

Bull SA, Allen VM, Domingue G, et al. (2006b) Sources of *Campylobacter spp.* colonizing housed broiler flocks during rearing. Appl Environ Microbiol 72:645–652. doi: 10.1128/AEM.72.1.645-652.2006

Bullitt E, Makowski L (1995) Structural polymorphism of bacterial adhesion pili. Nature 373:164–167. doi: 10.1038/373164a0

Button Gomez E (2008) Lessons learned from an elementary school norovirus outbreak. J Sch Nurs 24:388–97. doi: 10.1177/1059840508324069

Cadranel JF, Rufat P, Degos F (2000) Practices of liver biopsy in France: results of a prospective nationwide survey. For the Group of Epidemiology of the French Association for the Study of the Liver (AFEF). Hepatology 32:477–81. doi: 10.1053/jhep.2000.16602

Calderon-Margalit R, Sheffer R, Halperin T, et al. (2005) A large-scale gastroenteritis outbreak associated with Norovirus in nursing homes. Epidemiol Infect 133:35–40.

Calva JJ, Ruiz-Palacios GM, Lopez-Vidal AB, et al. (1988) Cohort study of intestinal infection with campylobacter in Mexican children. Lancet 331:503–506. doi: 10.1016/S0140-6736(88)91297-4

Campos PRA, Gordo I (2006) Pathogen genetic variation in small-world host contact structures. J Stat-Mech Theory 2006:L12003–L12003. doi: 10.1088/1742-5468/2006/12/L12003

Candotti D, Temple J, Sarkodie F, Allain J-P (2003) Frequent recovery and broad genotype 2 diversity characterize hepatitis C virus infection in Ghana, West Africa. J Virol 77:7914–23. doi: 10.1128/JVI.77.14.7914-7923.2003

Caprioli A, Morabito S, Brugère H, Oswald E (2005) Enterohaemorrhagic *Escherichia coli*: emerging issues on virulence and modes of transmission. Vet Res 36:289–311. doi: 10.1051/vetres:2005002

Carnieli P, de Novaes Oliveira R, Macedo CI, et al. (2011) Phylogeography of rabies virus isolated from dogs in Brazil between 1985 and 2006. Arch Virol 156:1007–12. doi: 10.1007/s00705-011-0942-y

Caul EO (1994) Small round structured viruses: airborne transmission and hospital control. Lancet 343:1240–2.

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis. Models and estimation procedures. Am J Hum Genet 19:233–57.

Centers for Disease Control and Prevention (1998) Recommendations for prevention and control of hepatitis C virus (HCV) infection and HCV-related chronic disease. Morb Mortal Wkly Rep 47:1–38.

Chamberlain RW, Adams N, Saeed AA, et al. (1997) Complete nucleotide sequence of a type 4 hepatitis C virus variant, the predominant genotype in the Middle East. J Gen Virol 78:1341–7.

Chan JZ-M, Pallen MJ, Oppenheim B, Constantinidou C (2012) Genome sequencing in clinical microbiology. Nature 30:1068–1071. doi: 10.1038/nbt.2410

Chan MCW, Chan PKS (2013) Complete genome sequence of a novel recombinant human norovirus genogroup II genotype 4 strain associated with an epidemic during summer of 2012 in Hong Kong. Genome Announc 1:e00140–12. doi: 10.1128/genomeA.00140-12

Check Hayden E (2012) Nanopore genome sequencer makes its debut. Nature. doi: 10.1038/nature.2012.10051

Chen FC, Li WH (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet 68:444–456. doi: 10.1086/318206

Chen SL, Morgan TR (2006) The natural history of hepatitis C virus (HCV) infection. Int J Med Sci 3:47–52.

Chin C-S, Sorenson J, Harris JB, et al. (2011) The origin of the Haitian cholera outbreak strain. N Engl J Med 364:33–42. doi: 10.1056/NEJMoa1012928

Cho J-C, Tiedje JM (2001) Bacterial species determination from DNA-DNA hybridization by using genome fragments and DNA microarrays. Appl Environ Microbiol. doi: 10.1128/AEM.67.8.3677-3682.2001

Choo QL, Kuo G, Weiner AJ, et al. (1989) Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. Science 244:359–62. doi: 10.1126/science.2523562

Clark CG, Price L, Ahmed R, et al. (2003) Characterization of waterborne outbreak-associated *Campylobacter jejuni*, Walkerton, Ontario. Emerg Infect Dis 9:1232–1241. doi: 10.3201/eid0910.020584

Cleaveland S, Laurenson MK, Taylor LH (2001) Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. Phil Trans R Soc B 356:991–9. doi: 10.1098/rstb.2001.0889

Coker AO, Isokpehi RD, Thomas BN, et al. (2002) Human campylobacteriosis in developing countries. Emerg Infect Dis 8:237–243. doi: 10.3201/eid0803.010233

Colin C, Lanoir D, Touzet S, et al. (2001) Sensitivity and specificity of third-generation hepatitis C virus antibody detection assays: an analysis of the literature. J Viral Hepatitis 8:87–95. doi: 10.1046/j.1365-2893.2001.00280.x

Colombo M, Oldani S, Donato MF, et al. (1987) A multicenter, prospective study of posttransfusion hepatitis in Milan. Hepatology 7:709–12. doi: 10.1002/hep.1840070415

Contreras AM, Ochoa-Jiménez RJ, Celis A, et al. (2010) High antibody level: an accurate serologic marker of viremia in asymptomatic people with hepatitis C infection. Transfusion 50:1335–43. doi: 10.1111/j.1537-2995.2009.02571.x

Cooke R, Goddard S, Golland J (2003) Costing a major hospital outbreak of gastroenteritis due to norovirus (Norwalk-like virus). Br J Infect Control 4:18–21. doi: 10.1177/175717740300400207

Cookson BD, Aparicio P, Deplano A, et al. (1996) Inter-centre comparison of pulsed-field gel electrophoresis for the typing of methicillin-resistant *Staphylococcus aureus*. J Med Microbiol 44:179–84.

Cooper BS, Medley GF, Bradley SJ, Scott GM (2008) An augmented data method for the analysis of nosocomial infection data. Am J Epidemiol 168:548–57. doi: 10.1093/aje/kwn176

Corbel MJ (1997) Brucellosis: an overview. Emerg Infect Dis 3:213–21. doi: 10.3201/eid0302.970219

Corry JEL, Atabay HI (2001) Poultry as a source of *Campylobacter* and related organisms. J Appl Microbiol 90:96S–114S. doi: 10.1046/j.1365-2672.2001.01358.x

Cottam EM, Wadsworth J, Shaw AE, et al. (2008) Transmission pathways of foot-and-mouth disease virus in the United Kingdom in 2007. PLoS Pathog 4:e1000050. doi: 10.1371/journal.ppat.1000050

Cramer N, Klockgether J, Wrasman K, et al. (2011) Microevolution of the major common *Pseudomonas aeruginosa* clones C and PA14 in cystic fibrosis lungs. Environ Microbiol 13:1690–704. doi: 10.1111/j.1462-2920.2011.02483.x

Cranston KA, Rannala B (2007) Summarizing a posterior distribution of trees using agreement subtrees. Syst Biol 56:578–90. doi: 10.1080/10635150701485091

Crick F (1970) Central dogma of molecular biology. Nature 227:561–563.

Crick FHC, Barnett L, Brenner S, Watts-Tobin RJ (1961) General nature of the genetic code for proteins. Nature 192:1227–32.

Criscuolo A, Berry V, Douzery EJP, Gascuel O (2006) SDM: a fast distance-based approach for (super) tree building in phylogenomics. Syst Biol 55:740–55. doi: 10.1080/10635150600969872

Cule M, Donnelly P Stochastic modelling and inference in electronic hospital databases for the spread of infections: *Clostridium difficile* transmission in Oxfordshire hospitals 2007-2010. Submitted to Ann. Appl. Stat.

Darwin C (1859) On the Origin of Species. 475:424. doi: 10.1038/475424a

Daszak P, Cunningham AA, Hyatt AD (2001) Anthropogenic environmental change and the emergence of infectious diseases in wildlife. Acta Trop 78:103–16. doi: 10.1016/S0001-706X(00)00179-0

Davey JW, Hohenlohe PA, Etter PD, et al. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510. doi: 10.1038/nrg3012

de la Rua-Domenech R (2006) Human *Mycobacterium bovis* infection in the United Kingdom: Incidence, risks, control measures and review of the zoonotic aspects of bovine tuberculosis. Tuberculosis 86:77–109. doi: 10.1016/j.tube.2005.05.002

de Wit MAS, Widdowson MA, Vennema H, et al. (2007) Large outbreak of norovirus: the baker who should have known better. J Infect 55:188–93. doi: 10.1016/j.jinf.2007.04.005

Dearlove B, Wilson DJ (2013) Coalescent inference for infectious disease: meta-analysis of hepatitis C. Phil Trans R Soc B 368:20120314. doi: 10.1098/rstb.2012.0314

Debbink K, Donaldson EF, Lindesmith LC, Baric RS (2012a) Genetic mapping of a highly variable norovirus GII.4 blockade epitope: potential role in escape from human herd immunity. J Virol 86:1214–26. doi: 10.1128/JVI.06189-11

Debbink K, Lindesmith LC, Donaldson EF, Baric RS (2012b) Norovirus immunity and the great escape. PLoS Pathog 8:e1002921. doi: 10.1371/journal.ppat.1002921

Degnan JH, Salter LA (2005) Gene tree distributions under the coalescent process. Evolution 59:24–37.

Dekeyser P, Gossuin-Detrain M, Butzler JP, Sternon J (1972) Acute enteritis due to related vibrio: first positive stool cultures. J Infect Dis 125:390–2. doi: 10.1093/infdis/125.4.390

Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6:361–75. doi: 10.1038/nrg1603

Didelot X (2013) Genomic analysis to improve the management of outbreaks of bacterial infection. Expert Rev Anti Infect Ther 11:335–7. doi: 10.1586/eri.13.15

Didelot X, Bowden R, Wilson DJ, et al. (2012a) Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet 13:601–12. doi: 10.1038/nrg3226

Didelot X, Falush D (2007) Inference of bacterial microevolution using multilocus sequence data. Genetics 175:1251–66. doi: 10.1534/genetics.106.063305

Didelot X, Méric G, Falush D, Darling AE (2012b) Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. BMC Genomics 13:256. doi: 10.1186/1471-2164-13-256

Dietrich MR, Skipper RA (2007) Manipulating underdetermination in scientific controversy: the case of the molecular clock. Perspect Sci 15:295–326.

Dingle KE, Colles FM, Falush D, Maiden MCJ (2005) Sequence Typing and Comparison of Population Biology of *Campylobacter coli* and *Campylobacter jejuni*. J Clin Microbiol 43:340–347. doi: 10.1128/JCM.43.1.340-347.2005

Dingle KE, Colles FM, Wareing DRA, et al. (2001) Multilocus sequence typing system for *Campylobacter jejuni*. J Clin Microbiol 39:14–23. doi: 10.1128/JCM.39.1.14-23.2001

Dingle KE, Norovirus Infection Control in Oxfordshire Communities Hospitals (2004) Mutation in a Lordsdale norovirus epidemic strain as a potential indicator of transmission routes. J Clin Microbiol 42:3950–3957. doi: 10.1128/JCM.42.9.3950

Dobell C (1932) Antony van Leeuwenhoek and his "Little Animals."Harcourt, Brace and Company, New York

Dobson AP, Carper ER (1996a) Infectious diseases and human population history. BioSc 46:115–126.

Dobson AP, Carper ER (1996b) Infectious diseases and human population history. Bioscience 46:115–126. doi: 10.2307/1312814

Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 36:e105. doi: 10.1093/nar/gkn425

Donaldson EF, Lindesmith LC, Lobue AD, Baric RS (2010) Viral shape-shifting: norovirus evasion of the human immune system. Nat Rev Microbiol 8:231–41. doi: 10.1038/nrmicro2296

Donnelly P, Tavaré S (1995) Coalescents and genealogical structure under neutrality. Annu Rev Genet 29:401–421. doi: 10.1146/annurev.ge.29.120195.002153

Dorner M, Horwitz JA, Robbins JB, et al. (2011) A genetically humanized mouse model for hepatitis C virus infection. Nature 474:208–11. doi: 10.1038/nature10168

Doyle JJ (1992) Gene trees and species trees: Molecular systematics as one-character taxonomy. Syst Bot 17:144–163. doi: 10.2307/2419070

Drummond A, Pybus OG, Rambaut A (2003a) Inference of viral evolutionary rates from molecular sequences. Adv Parasitol 54:331–58.

Drummond AJ (2010) Summarizing posterior trees. http://beast.bio.ed.ac.uk/Summarizing_posterior_trees. Accessed 29 Aug 2013

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. PLoS Biol 4:e88. doi: 10.1371/journal.pbio.0040088

Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161:1307–1320.

Drummond AJ, Pybus OG, Rambaut A, et al. (2003b) Measurably evolving populations. Trends Ecol Evol 18:481–488. doi: 10.1016/S0169-5347(03)00216-7

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7:214. doi: 10.1186/1471-2148-7-214

Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22:1185–92. doi: 10.1093/molbev/msi103

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969–73. doi: 10.1093/molbev/mss075

Duizer E, Bijkerk P, Rockx B, et al. (2004) Inactivation of caliciviruses. Appl Environ Microbiol 70:4538–4543. doi: 10.1128/AEM.70.8.4538

Duplantier J-M, Duchemin J-B, Chanteau S, Carniel E (2005) From the recent lessons of the Malagasy foci towards a global understanding of the factors involved in plague reemergence. Vet Res 36:437–453. doi: 10.1051/vetres:2005007

Eden J-S, Bull RA, Tu E, et al. (2010) Norovirus GII.4 variant 2006b caused epidemics of acute gastroenteritis in Australia during 2007 and 2008. J Clin Virol 49:265–271. doi: 10.1016/j.jcv.2010.09.001

Eden J-S, Tanaka MM, Boni MF, et al. (2013) Recombination within the pandemic norovirus GII.4 lineage. J Virol 87:6270–82. doi: 10.1128/JVI.03464-12

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–7. doi: 10.1093/nar/gkh340

Edgar RC, Batzoglou S (2006) Multiple sequence alignment. Curr Opin Struct Biol 16:368–73. doi: 10.1016/j.sbi.2006.04.004

Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7:1–26.

Efron B, Tibshirani RJ (1994) An Introduction to the Bootstrap. Chapman & Hall/CRC, New York

El-Zanaty F, Way A (2009) Egypt Demographic and Health Survey 2008. 252.

Engberg J, Aarestrup FM, Taylor DE, et al. (2001) Quinolone and macrolide resistance in *Campylobacter jejuni* and *C. coli*: resistance mechanisms and trends in human isolates. Emerg Infect Dis 7:24–34. doi: 10.3201/eid0701.700024

Engberg J, Gerner-Smidt P, Scheutz F, et al. (1998) Water-borne *Campylobacter jejuni* infection in a Danish town - a 6-week continuous source outbreak. Clin Microbiol Infect 4:648–656. doi: 10.1111/j.1469-0691.1998.tb00348.x

Enright MC, Spratt BG (2011) The genomic view of bacterial diversification. Science 331:407–9. doi: 10.1126/science.1201690

Escherich T (1886) Beitrage zur Kenntniss der Darmbacterien. III. Ueber das Vorkommen von Vibrionen im Darmcanal und den Stuhlgangen der Sauglinge. Münchener medizinische Wochenschrift 33:815–817.

Esteban JI, Sauleda S, Quer J (2008) The changing epidemiology of hepatitis C virus infection in Europe. J Hepatol 48:148–62. doi: 10.1016/j.jhep.2007.07.033

Eyre DW, Cule ML, Wilson DJ, et al. (2013) Diverse sources of *C. difficile* infection identified on whole-genome sequencing. N Engl J Med 369:1195–205. doi: 10.1056/NEJMoa1216064

Eyre DW, Golubchik T, Gordon NC, et al. (2012) A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and *Clostridium difficile* for outbreak detection and surveillance. BMJ Open. doi: 10.1136/bmjopen-2012-001124

Fankhauser RL, Monroe SS, Noel JS, et al. (2002) Epidemiologic and molecular trends of "Norwalk-like viruses" associated with outbreaks of gastroenteritis in the United States. J Infect Dis 186:1–7. doi: 10.1086/341085

Farr W (1852) Influence of elevation on the fatality of cholera. J Stat Soc London 15:155–183. doi: 10.2307/2338305

Fassio E (2010) Hepatitis C and hepatocellular carcinoma. Ann Hepatol 9:S119–S122.

Fauci AS (2001) Infectious diseases: considerations for the 21st century. Clin Infect Dis 32:675–85. doi: 10.1086/319235

Fearnhead P, Smith NGC, Barrigas M, et al. (2005) Analysis of recombination in *Campylobacter jejuni* from MLST population data. J Mol Evol 61:333–40. doi: 10.1007/s00239-004-0316-0

Fedurco M, Romieu A, Williams S, et al. (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res 34:e22. doi: 10.1093/nar/gnj023

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–76.

Felsenstein J (2004) Inferring Phylogenies. 38–44.

Felsenstein J (2005) PHYLIP (Phylogeny Inference Package). http://evolution.genetics.washington.edu/phylip.html. Accessed 31 Aug 2013

Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791.

Ferguson NM, Donnelly CA, Anderson RM (2001) The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. Science 292:1155–60. doi: 10.1126/science.1061020

Ferguson Wood EJ (1949) Classification of bacteria. Nature 164:867–9. doi: 10.1038/164867a0

Ferrero RL, Lee A (1988) Motility of *Campylobacter jejuni* in a viscous environment: comparison with conventional rod-shaped bacteria. J Gen Microbiol 134:53–9.

Fienberg SE (2006) When did Bayesian inference become "Bayesian"? Bayesian Anal 1:1–40. doi: 10.1214/06-BA101

Finney JM, Walker AS, Peto TEA, Wyllie DH (2011) An efficient record linkage scheme using graphical analysis for identifier error detection. BMC Med Inform Decis 11:7. doi: 10.1186/1472-6947-11-7

Fisher RA (1918) The correlation between relatives on the supposition of Mendalian inheritance. Philos Trans R Soc Edinburgh 52:399–433.

Fisher RA (1930) The Genetical Theory of Natural Selection. Clarendon, Oxford

Fleischmann RD, Adams MD, White O, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512.

Fonager J, Barzinci S, Fischer T (2013) Emergence of a new recombinant Sydney 2012 norovirus variant in Denmark, 26 December 2012 to 22 March 2013. Euro Surveill 18:1–6.

Forrester ML, Pettitt AN, Gibson GJ (2007) Bayesian inference of hospital-acquired infectious diseases and control measures given imperfect surveillance data. Biostatistics 8:383–401. doi: 10.1093/biostatistics/kxl017

Forsdyke DR (2006) Exons and introns. Evol. Bioinforma. Springer, New York, pp 207–224

Fouts DE, Mongodin EF, Mandrell RE, et al. (2005) Major structural differences and novel potential virulence mechanisms from the genomes of multiple campylobacter species. PLoS Biol 3:e15. doi: 10.1371/journal.pbio.0030015

Fraser C, Donnelly CA, Cauchemez S, et al. (2009) Pandemic potential of a strain of influenza A (H1N1): early findings. Science 324:1557–61. doi: 10.1126/science.1176062

Freeman VJ (1951) Studies on the virulence of bacteriophage-infected strains of *Corynebacterium diphtheriae*. J Bacteriol 61:675–88.

Fretz C, Jeannel D, Stuyver L, et al. (1995) HCV infection in a rural population of the Central African Republic (CAR): evidence for three additional subtypes of genotype 4. J Med Virol 137:435–437. doi: 10.1002/jmv.1890470423

Fried MW, Shiffman ML, Reddy KR, et al. (2002) Pegiterferon alfa-2a plus ribavirin for chronic hepatitis C virus infection. N Engl J Med 347:975–982. doi: 10.1056/NEJMoa020047

Friedman CR, Hoekstra RM, Samuel M, et al. (2004) Risk factors for sporadic *Campylobacter* infection in the United States: A case-control study in FoodNet sites. Clin Infect Dis 38:S285–96. doi: 10.1086/381598

Frost JA (2001) Current epidemiological issues in human campylobacteriosis. J Appl Microbiol 90:85S–95S. doi: 10.1046/j.1365-2672.2001.01357.x

Frost SDW, Volz EM (2010) Viral phylodynamics and the search for an "effective number of infections". Phil Trans R Soc B 365:1879–90. doi: 10.1098/rstb.2010.0060

Fu Y, Wang Y, Xia W, et al. (2011) New trends of HCV infection in China revealed by genetic analysis of viral sequences determined from first-time volunteer blood donors. J Viral Hepatitis 18:42–52. doi: 10.1111/j.1365-2893.2010.01280.x

Furuno M, Kasukawa T, Saito R, et al. (2003) CDS annotation in full-length cDNA sequence. Genome Res 13:1478–87. doi: 10.1101/gr.1060303

Fusco DN, Chung RT (2012) Novel therapies for hepatitis C: insights from the structure of the virus. Annu Rev Med 63:373–87. doi: 10.1146/annurev-med-042010-085715

Gallimore CI, Barreiros MAB, Brown DWG, et al. (2004a) Noroviruses associated with acute gastroenteritis in a children's day care facility in Rio de Janeiro, Brazil. Braz J Med Biol Res 37:321–326. doi: 10.1590/S0100-879X2004000300005

Gallimore CI, Cubitt D, du Plessis N, Gray JJ (2004b) Asymptomatic and symptomatic excretion of noroviruses during a hospital outbreak of gastroenteritis. J Clin Microbiol 42:9–13. doi: 10.1128/JCM.42.5.2271-2274.2004

Galton F (1897) The average contribution of several each ancestor to the heritage of the offspring. Proc R Soc 61:401–413.

Gane EJ, Stedman CA, Hyland RH, et al. (2013) Nucleotide polymerase inhibitor sofosbuvir plus ribavirin for hepatitis C. N Engl J Med 368:34–44. doi: 10.1056/NEJMoa1208953

Garnett GP, Holmes EC (1996) The ecology of emergent infectious disease. Bioscience 46:127–135. doi: 10.2307/1312815

Gaynes RP (2011) Germ theory: medical pioneers in infectious diseases. ASM Press, Washington DC

Gelberg L, Robertson MJ, Arangua L, et al. (2012) Prevalence, distribution, and correlates of hepatitis C virus infection among homeless adults in Los Angeles. Public Health Rep 127:407–21.

Geyer CJ (1991) Markov chain Monte Carlo maximum likelihood. In: Keramidas EM, Kaufman SM (eds) Comput. Sci. Stat. pp 156–163

Ghany MG, Strader DB, Thomas DL, Seeff LB (2009) Diagnosis, management, and treatment of hepatitis C: an update. Hepatology 49:1335–74. doi: 10.1002/hep.22759

Giammanco GM, De Grazia S, Tummolo F, et al. (2013) Norovirus variant GII.4/Sydney/2012 in Italy, winter 2012-2013. Emerg Infect Dis 19:1348–9. doi: 10.3201/eid1908.130619

Giesemann T, Egerer M, Jank T, Aktories K (2008) Processing of Clostridium difficile toxins. J Med Microbiol 57:690–6. doi: 10.1099/jmm.0.47742-0

Gilchrist MR (1998) Disease and infection in the American Civil War. Am Biol Teach 60:258–262. doi: 10.2307/4450468

Gilead (2013) Sovaldi prescribing information. 1–34.

Gillespie IA, O'Brien SJ, Frost JA, et al. (2006) Investigating vomiting and/or bloody diarrhoea in *Campylobacter jejuni* infection. J Med Microbiol 55:741–6. doi: 10.1099/jmm.0.46422-0

Gilliss D, Cronquist AB, Cartter M, et al. (2013) Incidence and trends of infection with pathogens transmitted commonly through food - Foodbourne Diseases Active Surveillance Network, 10 US Sites, 1996-2012. Morb Mortal Wkly Rep 62:283–287.

Giraud A, Matic I, Tenaillon O, et al. (2001) Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. Science 291:2606–8. doi: 10.1126/science.1056421

Glass RI, Stoll BJ, Huq MI, et al. (1983) Epidemiologic and clinical features of endemic *Campylobacter jejuni* infection in Bangladesh. J Infect Dis 148:292–6. doi: 10.1093/infdis/148.2.292

Goedert JJ, Chen BE, Preiss L, et al. (2007) Reconstruction of the hepatitis C virus epidemic in the US hemophilia population, 1940-1990. Am J Epidemiol 165:1443–53. doi: 10.1093/aje/kwm030

Goering R V (2010) Pulsed field gel electrophoresis: a review of application and interpretation in the molecular epidemiology of infectious disease. Infect Genet Evol 10:866–75. doi: 10.1016/j.meegid.2010.07.023

Gordo I, Gomes MGM, Reis DG, Campos PRA (2009) Genetic diversity in the SIR model of pathogen evolution. PLoS One 4:e4876. doi: 10.1371/journal.pone.0004876

Gould SJ (1986) Evolution and the triumph of homology, or why history matters. Am Sci 74:60–69.

Gram HC (1884) Über die isolierte Färbung der Schizomyceten in Schnitt- und Trockenpräparaten. Fortschr. Med., Vol 2. pp 185–89

Gray RR, Parker J, Lemey P, et al. (2011) The mode and tempo of hepatitis C virus evolution within and among hosts. BMC Evol Biol 11:131. doi: 10.1186/1471-2148-11-131

Green SM, Lambden PR, Caul EO, et al. (1995) Capsid diversity in small round-structured viruses: molecular characterization of an antigenically distinct human enteric calicivirus. Virus Res 37:271–283.

Greig JD, Lee MB (2009) Enteric outbreaks in long-term care facilities and recommendations for prevention: a review. Epidemiol Infect 137:145–55. doi: 10.1017/S0950268808000757

Grenfell BT, Pybus OG, Gog JR, et al. (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303:327–32. doi: 10.1126/science.1090727

Griekspoor P, Colles FM, McCarthy ND, et al. (2013) Marked host specificity and lack of phylogeographic population structure of *Campylobacter jejuni* in wild birds. Mol Ecol 22:1463–72. doi: 10.1111/mec.12144

Griffiths PL, Park RW (1990) Campylobacters associated with human diarrhoeal disease. J Appl Bacteriol 69:281–301. doi: 10.1111/j.1365-2672.1990.tb01519.x

Griffiths RC, Marjoram P (1996) Ancestral inference from samples of DNA sequences with recombination. J Comput Biol 3:479–502.

Griffiths RC, Tavaré S (1994) Sampling theory for neutral alleles in a varying environment. Phil Trans R Soc B 344:403–410. doi: 10.1098/rstb.1994.0079

Gripp E, Hlahla D, Didelot X, et al. (2011) Closely related *Campylobacter jejuni* strains from different sources reveal a generalist rather than a specialist lifestyle. BMC Genomics 12:584. doi: 10.1186/1471-2164-12-584

Guindon S, Dufayard J-F, Lefort V, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–21. doi: 10.1093/sysbio/syq010

Gundogdu O, Bentley SD, Holden MT, et al. (2007) Re-annotation and re-analysis of the *Campylobacter jejuni* NCTC11168 genome sequence. BMC Genomics. doi: 10.1186/1471-2164-8-162

Haeckel E (1866) Generelle Morphologie der Organismen: Allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte Descendenz-Theorie, Volume 2. Georg Reimer, Berlin

Haldane JBS (1932) The Causes of Evolution. Longmans, Green and Co. Ltd., London

Haldane JBS (1924) A mathematical theory of natural and artificial selection. Part I. Trans Cambridge Philos Soc 23:3–41.

Haldane JBS (1927) A mathematical theory of Natural and Artificial Selection. Part IV. Math Proc Cambridge Philos Soc 23:607–15. doi: 10.1017/S0305004100011750

Hall AJ, Lopman BA, Payne DC, et al. (2013) Norovirus Disease in the United States. Emerg Infect Dis 19:1198–1205. doi: 10.3201/eid1908.130465

Harbeck M, Seifert L, Hänsch S, et al. (2013) *Yersinia pestis* DNA from skeletal remains from the 6th century AD reveals insights into Justinianic Plague. PLoS Pathog 9:e1003349. doi: 10.1371/journal.ppat.1003349

Hardy ME (2005) Norovirus protein structure and function. FEMS Microbiol Lett 253:1–8. doi: 10.1016/j.femsle.2005.08.031

Harrington CS, Thomson-Carter FM, Carter PE (1997) Evidence for recombination in the flagellin locus of *Campylobacter jejuni*: implications for the flagellin gene typing scheme. J Clin Microbiol 35:2386–2392.

Harris N V, Weiss NS, Nolan CM (1986) The role of poultry and meats in the etiology of *Campylobacter jejuni/coli* enteritis. Am J Public Health 76:407–11. doi: 10.2105/AJPH.76.4.407

Harrison WA, Griffith CJ, Tennant D, Peters AC (2001) Incidence of *Campylobacter* and *Salmonella* isolated from retail chicken and associated packaging in South Wales. Lett Appl Microbiol 33:450–4. doi: 10.1046/j.1472-765X.2001.01031.x

Hasegawa M, Kishino H, Yano T (1985) Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA. J Mol Evol 22:160–174. doi: 10.1007/BF02101694

Hasegawa M, Kishino H, Yano T (1989) Estimation of branching dates among primates by molecular clocks of nuclear DNA which slowed down in Hominoidea. J Hum Evol 18:461–476.

Haß J, Matuszewski S, Cieslik D, Haase M (2011) The role of swine as "mixing vessel" for interspecies transmission of the influenza A subtype H1N1: a simultaneous Bayesian inference of phylogeny and ancestral hosts. Infect Genet Evol 11:437–41. doi: 10.1016/j.meegid.2010.12.001

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109. doi: 10.2307/2334940

Hauri AM, Armstrong GL, Hutin YJF (2004) The global burden of disease attributable to contaminated injections given in health care settings. Int J STD AIDS 15:7–16. doi: 10.1258/095646204322637182

Hazeleger WC, Wouters JA, Rombouts FM, Abee T (1998) Physiological activity of *Campylobacter jejuni* far below the minimal growth temperature. Appl Environ Microbiol 64:3917–3922.

Health Protection Agency (2013) Hospital norovirus outbreaks (England and Wales, weeks 49- 52/2012) and seasonal comparisons of recent years' norovirus laboratory reports. Health Protection Report: Weekly Report 7:8–9.

Health Protection Agency (2012) Laboratory reported *Campylobacter sp* infections reported by month, England and Wales, 2000-2011. http://www.hpa.org.uk/Topics/InfectiousDiseases/InfectionsAZ/Campylobacter/EpidemiologicalData/campyDataEwMonth1989to2008/. Accessed 14 Mar 2013

Hedge J, Lycett SJ, Rambaut A (2013) Real-time characterization of the molecular epidemiology of an influenza pandemic. Biol Lett 9:20130331. doi: 10.1098/rsbl.2013.0331

Hein J, Schierup MH, Wiuf C (2005) Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory. Oxford University Press, Oxford

Henderson DA (1976) The eradication of smallpox. Sci Am 235:25–33. doi: 10.1038/scientificamerican1076-25

Hennig W (1950) Grundzüge einer Theorie der Phylogenetischen Systematik. Dt. Zentralverl.

Hennig W (1966) Phylogenetic Systematics. University of Illinois Press, Urbana

Henquell C, Guglielmini J, Verbeeck J, et al. (2011) Evolutionary history of hepatitis C virus genotype 5a in France, a multicenter ANRS study. Infect Genet Evol 11:496–503. doi: 10.1016/j.meegid.2010.12.015

Hethcote HW (2000) The Mathematics of Infectious Diseases. SIAM Rev 42:599–653. doi: 10.2307/2653135

Higgins D, Lemey P (2009) Multiple sequence alignment. In: Lemey P, Salemi M, Vandamme A-M (eds) The Phylogenetic Handbook, 2nd ed. pp 68–99

Holmes EC (2010) The RNA virus quasispecies: fact or fiction? J Mol Biol 400:271–3. doi: 10.1016/j.jmb.2010.05.032

Hoofnagle JH (1997) Hepatitis C: the clinical spectrum of the disease. J Hepatol 26:15S–20S. doi: 10.1002/hep.510260703

Huang P, Farkas T, Marionneau S, et al. (2003) Noroviruses bind to human ABO, Lewis, and secretor histo-blood group antigens: identification of 4 distinct strain-specific patterns. J Infect Dis 188:19–31. doi: 10.1086/375742

Huang P, Farkas T, Zhong W, et al. (2005) Norovirus and histo-blood group antigens: demonstration of a wide spectrum of strain specificities and classification of two major binding groups among multiple binding patterns. J Virol 79:6714–6722. doi: 10.1128/JVI.79.11.6714

Hudson RR, Kaplan NL (1988) The coalescent process in models with selection and recombination. Genetics 120:831–40.

Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147–64.

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–5.

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254–67. doi: 10.1093/molbev/msj030

Hutin YJF, Hauri AM, Armstrong GL (2003) Use of injections in healthcare settings worldwide, 2000: literature review and regional estimates. BMJ. doi: http://dx.doi.org/10.1136/bmj.327.7423.1075

Hutson AM, Atmar RL, Estes MK (2004) Norovirus disease: changing epidemiology and host susceptibility factors. Trends Microbiol 12:279–87. doi: 10.1016/j.tim.2004.04.005

Hutson AM, Atmar RL, Graham DY, Estes MK (2002) Norwalk virus infection and disease is associated with ABO histo-blood group type. J Infect Dis 185:1335–7. doi: 10.1086/339883

Huttner B, Cordey S, Sauvan V, et al. (2013) O073: An outbreak of norovirus strain GII.4 Sydney in a geriatric teaching hospital. Antimicrob Resist Infect Control 2:O73. doi: 10.1186/2047-2994-2-S1-O73

Huxley J (1942) Evolution: The Modern Synthesis. Allen & Unwin, London

Iqbal Z, Caccamo M, Turner I, et al. (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet 44:226–32. doi: 10.1038/ng.1028

Ivanowski D (1892) Über die Mosaikkrankheit der Tabakspflanze. Bull Sci publié par l'Académie Impériale des Sci Saint-pétersbg 35:67–70.

Jacobson IM, McHutchison JG, Dusheiko G, et al. (2011) Telaprevir for previously untreated chronic hepatitis C virus infection. N Engl J Med 364:2405–16. doi: 10.1056/NEJMoa1012912

Jafari S, Copes R, Baharlou S, et al. (2010) Tattooing and the risk of transmission of hepatitis C: a systematic review and meta-analysis. Int J Infect Dis 14:e928–40. doi: 10.1016/j.ijid.2010.03.019

Jatapai A, Nelson KE, Chuenchitra T, et al. (2010) Prevalence and risk factors for hepatitis C virus infection among young Thai men. Am J Trop Med Hyg 83:433–9. doi: 10.4269/ajtmh.2010.09-0749

Jeannel D, Fretz C, Traore Y, et al. (1998) Evidence for high genetic diversity and long-term endemicity of hepatitis C virus genotypes 1 and 2 in West Africa. J Med Virol 55:92–7. doi: 10.1002/(SICI)1096-9071(199806)55:2<92::AID-JMV2>3.0.CO;2-I

Jeffreys H (1961) Theory of Probability, Third Edit. Oxford Clarendon Press, Oxford

Jenkins GM, Rambaut A, Pybus OG, Holmes EC (2002) Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. J Mol Evol 54:156–65. doi: 10.1007/s00239-001-0064-3

Jerome JP, Klahn BD, Bell JA, et al. (2012) Draft genome sequences of two Campylobacter jejuni clinical isolates, NW and D2600. J Bacteriol 194:5707–8. doi: 10.1128/JB.01338-12

Jiang X, Graham DY, Wang K, Estes MK (1990) Norwalk virus genome cloning and characterization. Science 250:1580–3.

Jiang X, Wang J, Graham DY, Estes MK (1992) Detection of Norwalk virus in stool by polymerase chain reaction. J Clin Microbiol 30:2529–34.

Johnson DL, Mead KR, Lynch RA, Hirst DVL (2013) Lifting the lid on toilet plume aerosol: a literature review with suggestions for future research. Am J Infect Control 41:254–8. doi: 10.1016/j.ajic.2012.04.330

Johnson NPAS, Mueller J (2002) Updating the accounts: global mortality of the 1918-1920 "Spanish" influenza pandemic. Bull Hist Med 76:105–115. doi: 10.1353/bhm.2002.0022

Johnson PC, Mathewson JJ, DuPont HL, Greenberg HB (1990) Multiple-challenge study of host susceptibility to Norwalk gastroenteritis in US adults. J Infect Dis 161:18–21.

Jolley KA, Maiden MCJ (2010) BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11:595. doi: 10.1186/1471-2105-11-595

Jombart T, Cori A, Didelot X, et al. (2014) Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. PLoS Comput Biol 10:e1003457. doi: 10.1371/journal.pcbi.1003457

Jones EL, Kramer A, Gaither M, Charles P (2007) Role of fomite contamination during an outbreak of norovirus on houseboats. Int J Environ Health Res 17:123–131.

Jones K (2001) Campylobacters in water, sewage and the environment. J Appl Microbiol 90:68S–79S. doi: 10.1046/j.1365-2672.2001.01355.x

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HH (ed) Mammalian Protein Metabolism, Vol III. Academic Press, New York, pp 21–132

Kapikian AZ, Wyatt RG, Dolin R, et al. (1972) Visualization by immune electron microscopy of a 27-nm particle associated with acute infectious nonbacterial gastroenteritis. J Virol 10:1075–1081.

Kaplan JE, Feldman R, Campbell DS, et al. (1982a) The frequency of a Norwalk-like pattern of illness in outbreaks of acute gastroenteritis. Am J Public Health 72:1329–32.

Kaplan JE, Gary GW, Baron RC, et al. (1982b) Epidemiology of Norwalk gastroenteritis and the role of Norwalk virus in outbreaks of acute nonbacterial gastroenteritis. Ann Intern Med 96:756–761. doi: 10.7326/0003-4819-96-6-756

Kapperud G, Espeland G, Wahl E, et al. (2003) Factors associated with increased and decreased risk of *Campylobacter* infection: a prospective case-control study in Norway. Am J Epidemiol 158:234–242. doi: 10.1093/aje/kwg139

Kapperud G, Lassen J, Ostroff SM, Aasen S (1992) Clinical features of sporadic *Campylobacter* infections in Norway. Scand J Infect Dis 24:741–9.

Karst SM, Wobus CE, Lay M, et al. (2003) STAT1-dependent innate immunity to a Norwalk-like virus. Science 299:1575–8. doi: 10.1126/science.1077905

Kass RE, Raftery AE (1995) Bayes Factors. J Am Stat Assoc 90:773–795.

Kearse M, Moir R, Wilson A, et al. (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–9. doi: 10.1093/bioinformatics/bts199

Kermack WO, McKendrick AG (1927) A Contribution to the Mathematical Theory of Epidemics. Proc R Soc A 115:700–721. doi: 10.1098/rspa.1927.0118

Khanna N, Goldenberger D, Graber P, et al. (2003) Gastroenteritis outbreak with norovirus in a Swiss university hospital with a newly identified virus strain. J Hosp Infect 55:131–136. doi: 10.1016/S0195-6701(03)00257-3

Kim HP, Crockett SD, Shaheen NJ (2013) The burden of gastrointestinal and liver disease around the world. In: Talley NJ, Locke GR, Moayyedi P, et al. (eds) GI Epidemiol. Dis. Clin. Methodol., 2nd ed. Wiley-Blackwell, pp 3–14

Kim WR (2002) The burden of hepatitis C in the United States. Hepatology 36:S30–S34. doi: 10.1002/hep.1840360705

Kim WR, Brown RS, Terrault NA, El-Serag H (2002) Burden of liver disease in the United States: summary of a workshop. Hepatology 36:227–42. doi: 10.1053/jhep.2002.34734

Kimura M (1955) Stochastic processes and distribution of gene frequencies under natural selection. Cold Spring Harb Symp Quant Biol 20:33–53.

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–6. doi: 10.1038/217624a0

Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61:893–903.

Kimura M (1986) DNA and the Neutral Theory. Phil Trans R Soc B 312:343–354. doi: 10.1098/rstb.1986.0012

Kimura M (1983) The Neutral Theory of Molecular Evolution. neutral theory Mol Evol. doi: citeulike-article-id:4441469

Kimura M (1980) A simple method for estimation evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120.

Kinde H, Genigeorgis CA, Pappaioanou M (1983) Prevalence of *Campylobacter jejuni* in chicken wings. Appl Environ Microbiol 45:1116–8.

King EO (1957) Human infections with *Vibrio fetus* and a closely related vibrio. J Infect Dis 101:119–128. doi: 10.1093/infdis/101.2.119

King EO (1962) The laboratory recognition of *Vibrio fetus* and a closely related *Vibrio* isolated from cases of human vibrosis. Ann NY Acad Sci 98:700–711. doi: 10.1111/j.1749-6632.1962.tb30591.x

Kingman JFC (1982a) On the Genealogy of Large Populations. J Appl Probab 19, Essays:27–43.

Kingman JFC (1982b) The coalescent. Stoch Proc Appl 13:235–248. doi: 10.1016/0304-4149(82)90011-4

Kircher M, Stenzel U, Kelso J (2009) Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome Biol 10:R83. doi: 10.1186/gb-2009-10-8-r83

Koch R (1878) Untersuchungen über die Aetiologie der Wundinfectionskrankheiten. Verlag von F. C. W. Vogel, Leipzig

Koch R (1891) Über bakteriologische Forschung. Vortrag in der 1. allgemeinen Sitzung des X. Internationalen Medicinischen Congresses am 4. August 1890

Koelle K, Rasmussen DA (2012) Rates of coalescence for common epidemiological models at equilibrium. J R Soc Interface 9:997–1007. doi: 10.1098/rsif.2011.0495

Koopmans M (2008) Progress in understanding norovirus epidemiology. Curr Opin Infect Dis 21:544–52. doi: 10.1097/QCO.0b013e3283108965

Köser CU, Holden MTG, Ellington MJ, et al. (2012) Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N Engl J Med 366:2267–75. doi: 10.1056/NEJMoa1109910

Kroneman A, Vennema H, Deforche K, et al. (2011) An automated genotyping tool for enteroviruses and noroviruses. J Clin Virol 51:121–5. doi: 10.1016/j.jcv.2011.03.006

Kroneman A, Verhoef L, Harris J, et al. (2008) Analysis of integrated virological and epidemiological reports of norovirus outbreaks collected within the foodborne viruses in Europe network from 1 July 2001 to 30 June 2006. J Clin Microbiol 46:2959–2965. doi: 10.1128/JCM.00499-08

Kuhner MK (2006) LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics 22:768–70. doi: 10.1093/bioinformatics/btk051

Kuhner MK, Felsenstein J (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol 11:459–68.

Kühnert D, Stadler T, Vaughan TG, Drummond AJ (2014) Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth − death SIR model. J R Soc Interface 11:20131196. doi: 10.1098/rsif.2013.1106

Kumar S (2005) Molecular clocks: four decades of evolution. Nat Rev Genet 6:654–62. doi: 10.1038/nrg1659

Kundu S, Lockwood J, Depledge DP, et al. (2013) Next-generation whole genome sequencing identifies the direction of norovirus transmission in linked patients. Clin Infect Dis 57:407–14. doi: 10.1093/cid/cit287

Kurbanov F, Tanaka Y, Elkady a, et al. (2007) Tracing hepatitis C and Delta viruses to estimate their contribution in HCC rates in Mongolia. J Viral Hepatitis 14:667–74. doi: 10.1111/j.1365-2893.2007.00864.x

La Scola B, Audic S, Robert C, et al. (2003) A giant virus in amoebae. Science 299:2033. doi: 10.1126/science.1081867

Lam N-CV, Gotsch PB, Langan RC (2010) Caring for pregnant women and newborns with hepatitis B or C. Am Fam Physician 82:1225–9.

Lamarck JBPA (1809) Philosophie zoologique, ou exposition des considérations relatives à l'histoire naturelle des animaux. Tome second. 235.

Lamarre D, Anderson PC, Bailey M, et al. (2003) An NS3 protease inhibitor with antiviral effects in humans infected with hepatitis C virus. Nature 426:186–9. doi: 10.1038/nature02099

Langmuir AD (1976) William Farr: founder of modern concepts of surveillance. Int J Epidemiol 5:13–18. doi: 10.1093/ije/5.1.13

Lankester ER (1870) On the use of term homology in modern zoology, and the distinction between homogenetic and homoplasic agreements. Ann Mag Nat Hist 6:34–43.

Lavanchy D (2011) Evolving epidemiology of hepatitis C virus. Clin Microbiol Infect 17:107–15. doi: 10.1111/j.1469-0691.2010.03432.x

Lawitz E, Mangia A, Wyles D, et al. (2013) Sofosbuvir for previously untreated chronic hepatitis C infection. N Engl J Med 368:1878–87. doi: 10.1056/NEJMoa1214853

Lay MK, Atmar RL, Guix S, et al. (2010) Norwalk virus does not replicate in human macrophages or dendritic cells derived from the peripheral blood of susceptible humans. Virology 406:1–11. doi: 10.1016/j.virol.2010.07.001

Lefébure T, Pavinski Bitar PD, Suzuki H, Stanhope MJ (2010) Evolutionary dynamics of complete *Campylobacter* pan-genomes and the bacterial species concept. Genome Biol Evol 2:646–55. doi: 10.1093/gbe/evq048

Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009) Bayesian phylogeography finds its roots. PLoS Comput Biol 5:e1000520. doi: 10.1371/journal.pcbi.1000520

Leung WK, Chan PKS, Lee NLS, Sung JJY (2010) Development of an in vitro cell culture model for human noroviruses and its clinical application. Hong Kong Med J 16:18–21.

Leventhal GE, Kouyos R, Stadler T, et al. (2012) Inferring epidemic contact structure from phylogenetic trees. PLoS Comput Biol 8:e1002413. doi: 10.1371/journal.pcbi.1002413

Levins R (1968a) Evolution in changing environments; some theoretical explorations. Monographs in Population Biology, no. 2. Princeton University Press, Princeton, NJ

Levins R (1968b) Some demographic and genetic consequences of environmental heterogeneity for biological control. Bull Entomol Soc Am 15:237–249.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–60. doi: 10.1093/bioinformatics/btp324

Li WLS, Drummond AJ (2012) Model averaging and Bayes factor calculation of relaxed molecular clocks in Bayesian phylogenetics. Mol Biol Evol 29:751–61. doi: 10.1093/molbev/msr232

Lieberman TD, Michel J-B, Aingaran M, et al. (2011) Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. Nat Genet 43:1275–80. doi: 10.1038/ng.997

Lin C, Kwong AD, Perni RB (2006) Discovery and development of VX-950, a novel, covalent, and reversible inhibitor of hepatitis C virus NS3.4A serine protease. Infect Disord - drug targets 6:3–16. doi: 10.2174/187152606776056706

Lindenbach BD, Evans MJ, Syder AJ, et al. (2005) Complete replication of hepatitis C virus in cell culture. Science 309:623–6. doi: 10.1126/science.1114016

Lindesmith L, Moe C, Marionneau S, et al. (2003) Human susceptibility and resistance to Norwalk virus infection. Nat Med 9:548–553. doi: 10.1038/nm860

Lindesmith LC, Donaldson EF, Baric RS (2011) Norovirus GII.4 strain antigenic variation. J Virol 85:231–42. doi: 10.1128/JVI.01364-10

Lindesmith LC, Donaldson EF, Lobue AD, et al. (2008) Mechanisms of GII.4 norovirus persistence in human populations. PLoS Med 5:e31. doi: 10.1371/journal.pmed.0050031

Linhartová I, Bumba L, Mašín J, et al. (2010) RTX proteins: a highly diverse family secreted by a common mechanism. FEMS Microbiol Rev 34:1076–112. doi: 10.1111/j.1574-6976.2010.00231.x

Linnaeus C (1758) Systema naturae per regna tria naturae, secundim classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis. Vol. 10. Laurentius Salvius, Stockholm

Lipsitch M, Cohen T, Cooper B, et al. (2003) Transmission dynamics and control of severe acute respiratory syndrome. Science 300:1966–70. doi: 10.1126/science.1086616

Little LK (2007) Plague and the End of Antiquity: The Pandemic of 541-750. 3–32.

Liu L (2008) BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24:2542–3. doi: 10.1093/bioinformatics/btn484

Liu P, Yuen Y, Hsiao H-M, et al. (2010) Effectiveness of liquid soap and hand sanitizer against Norwalk virus on contaminated hands. Appl Environ Microbiol 76:394–399. doi: 10.1128/AEM.01729-09

Loman NJ, Constantinidou C, Chan JZM, et al. (2012) High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat Rev Microbiol 10:599–606. doi: 10.1038/nrmicro2850

Lopman B, Andrews N, Sarangi J, et al. (2005) Institutional risk factors for outbreaks of nosocomial gastroenteritis: survival analysis of a cohort of hospital units in South-west England, 2002 – 2003. J Hosp Infect 60:135–143. doi: 10.1016/j.jhin.2004.10.021

Lopman BA, Reacher M, Gallimore C, et al. (2003) A summertime peak of "winter vomiting disease": Surveillance of noroviruses in England and Wales, 1995 to 2002. BMC Public Health. doi: 10.1186/1471-2458-3-13

Louis VR, Gillespie IA, O'Brien SJ, et al. (2005) Temperature-driven *Campylobacter* seasonality in England and Wales. Appl Environ Microbiol 71:85–92. doi: 10.1128/AEM.71.1.85-92.2005

Lu L, Nakano T, He Y, et al. (2005) Hepatitis C virus genotype distribution in China: predominance of closely related subtype 1b isolates and existence of new genotype 6 variants. J Med Virol 75:538–49. doi: 10.1002/jmv.20307

Lund M, Nordentoft S, Pedersen K, Madsen M (2004) Detection of *Campylobacter* spp. in chicken fecal samples by real-time PCR. J Clin Microbiol 42:5125–5132. doi: 10.1128/JCM.42.11.5125–5132.2004

Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21:936–9. doi: 10.1101/gr.111120.110

Lycett S, McLeish NJ, Robertson C, et al. (2012a) Origin and fate of A/H1N1 influenza in Scotland during 2009. J Gen Virol 93:1253–60. doi: 10.1099/vir.0.039370-0

Lycett SJ, Baillie G, Coulter E, et al. (2012b) Estimating reassortment rates in co-circulating Eurasian swine influenza viruses. J Gen Virol 93:2326–36. doi: 10.1099/vir.0.044503-0

Maddison WP (1997) Gene trees in species trees. Syst Biol 46:523–536. doi: 10.1093/sysbio/46.3.523

Magiorkinis G, Magiorkinis E, Paraskevis D, et al. (2009) The global spread of hepatitis C virus 1a and 1b: a phylodynamic and phylogeographic analysis. PLoS Med 6:e1000198. doi: 10.1371/journal.pmed.1000198

Mai H, Jin M, Guo X, et al. (2013) Clinical and epidemiologic characteristics of norovirus GII.4 Sydney during winter 2012-13 in Beijing, China following its global emergence. PLoS One 8:e71483. doi: 10.1371/journal.pone.0071483

Maiden MCJ (2006) Multilocus sequence typing of bacteria. Annu Rev Microbiol 60:561–88. doi: 10.1146/annurev.micro.59.030804.121325

Maiden MCJ, Bygraves JA, Feil E, et al. (1998) Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA 95:3140–3145.

Manns MP, McHutchison JG, Gordon SC, et al. (2001) Peginterferon alfa-2b plus ribavirin compared with interferon alfa-2b plus ribavirin for initial treatment of chronic hepatitis C: a randomised trial. Lancet 358:958–65.

Margush T, McNorris FR (1981) Consensus n-trees. Bull Math Biol 43:239–244. doi: 10.1016/S0092-8240(81)90019-7

Maritschnik S, Kanitz EE, Simons E, et al. (2013) A food handler-associated, foodborne norovirus GII.4 Sydney 2012-outbreak following a wedding dinner, Austria, October 2012. Food Environ Virol. doi: 10.1007/s12560-013-9127-z

Markov AA (1971) Extension of the limit theorems of probability theory to a sum of variables connected in a chain. In: Howard R (ed) Dynamic Probabilistic Systems, volume 1: Markov Chains. John Wiley and Sons, p 551

Martell M, Esteban J, Quer J, et al. (1992) Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution. J Virol 66:3225–3229.

Martella V, Campolo M, Lorusso E, et al. (2007) Norovirus in captive lion cub (*Panthera leo*). Emerg Infect Dis 13:1071–1073. doi: 10.3201/eid1307.070268

Martella V, Lorusso E, Decaro N, et al. (2008) Detection and molecular characterization of a canine norovirus. Emerg Infect Dis 14:1306–8. doi: 10.3201/eid1408.080062

Martella V, Medici MC, De Grazia S, et al. (2013) Evidence for recombination between pandemic GII.4 norovirus strains New Orleans 2009 and Sydney 2012. J Clin Microbiol 51:3855–7. doi: 10.1128/JCM.01847-13

Matheï C, Van Dooren S, Lemey P, et al. (2008) The epidemic history of hepatitis C among injecting drug users in Flanders, Belgium. J Viral Hepatitis 15:399–408. doi: 10.1111/j.1365-2893.2007.00950.x

Mathijs E, Denayer S, Palmeira L, et al. (2011) Novel norovirus recombinants and GII.4 sub-lineages associated with outbreaks between 2006 and 2010 in Belgium. Virol J 8:310. doi: 10.1186/1743-422X-8-310

May RM, Nowak MA (1994) Superinfection, metapopulation dynamics, and the evolution of diversity. J Theor Biol 170:95–114. doi: 10.1006/jtbi.1994.1171

Maynard Smith J, Smith NH (1998) Detecting recombination from gene trees. Mol Biol Evol 15:590–9.

Mayr E (1982) The Growth of Biological Thought: Diversity, Evolution and Inheritence. Belknap Press, Cambridge, Mass.

McAdam PR, Templeton KE, Edwards GF, et al. (2012) Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant <i>Staphylococcus aureus<\i>. Proc Natl Acad Sci USA 109:9107–12. doi: 10.1073/pnas.1202869109

McCarthy M, Estes MK, Hyams KC (2000) Norwalk-like virus infection in military forces: epidemic potential, sporadic disease, and the future direction of prevention and control efforts. J Infect Dis 181 Suppl:S387–91. doi: 10.1086/315582

McCarthy ND, Colles FM, Dingle KE, et al. (2007) Host-associated genetic import in *Campylobacter jejuni*. Emerg Infect Dis 13:267–72. doi: 10.3201/eid1302.060620

McFadyean J, Stockman S (1913) Report of the Departmental Committee appointed by the Board of Agriculture and Fisheries to inquire into Epizootic Abortion. III. Abortion in Sheep. London

McHutchison JG, Everson GT, Gordon SC, et al. (2009) Telaprevir with peginterferon and ribavirin for chronic HCV genotype 1 infection. N Engl J Med 360:1827–38. doi: 10.1056/NEJMoa0806104

McHutchison JG, Manns MP, Muir AJ, et al. (2010) Telaprevir for previously treated chronic HCV infection. N Engl J Med 362:1292–303. doi: 10.1056/NEJMoa0908014

Mead PS, Slutsker L, Dietz V, et al. (1999) Food-related illness and death in the United States. Emerg Infect Dis 5:607–25. doi: 10.3201/eid0506.990624

Medini D, Serruto D, Parkhill J, et al. (2008) Microbiology in the post-genomic era. Nat Rev Microbiol 6:419–30. doi: 10.1038/nrmicro1901

Meldrum RJ, Griffiths JK, Smith RMM, Evans MR (2005) The seasonality of human campylobacter infection and *Campylobacter* isolates from fresh, retail chicken in Wales. Epidemiol Infect 133:49–52. doi: 10.1017/S0950268804003188

Mendelson JA (2002) "Like all that lives": biology, medicine and bacteria in the age of Pasteur and Koch. Hist Philos Life Sci 24:22–23.

Mengarelli S, Correa G, Farias A, et al. (2006) ¿Por què el virus C en Cruz del Eje? Acta Gastroenterol Latinoam 36:68.

Mercer DF, Schiller DE, Elliott JF, et al. (2001) Hepatitis C virus replication in mice with chimeric human livers. Nat Med 7:927–33. doi: 10.1038/90968

Metropolis N, Rosenbluth AW, Rosenbluth MN, et al. (1953) Equation of state calculations by fast computing machines. J Chem Phys 21:1087–92. doi: 10.1063/1.1699114

Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11:31–46. doi: 10.1038/nrg2626

Minin VN, Suchard MA (2008a) Counting labeled transitions in continuous-time Markov models of evolution. J Math Biol 56:391–412. doi: 10.1007/s00285-007-0120-8

Minin VN, Suchard MA (2008b) Fast, accurate and simulation-free stochastic mapping. Phil Trans R Soc B 363:3985–95. doi: 10.1098/rstb.2008.0176

Moore JE, Barton MD, Blair IS, et al. (2006) The epidemiology of antibiotic resistance in *Campylobacter*. Microbes Infect 8:1955–66. doi: 10.1016/j.micinf.2005.12.030

Moore JE, Corcoran D, Dooley JSG, et al. (2005) *Campylobacter*. Vet Res 36:351–382. doi: 10.1051/vetres:2005012

Moradpour D, Penin F, Rice CM (2007) Replication of hepatitis C virus. Nat Rev Microbiol 5:453–63. doi: 10.1038/nrmicro1645

Moran PAP (1957) Random processes in genetics. Proc Cambridge Philos Soc 54:60–71.

Moreira D, López-García P (2009) Ten reasons to exclude viruses from the tree of life. Nat Rev Microbiol 7:306–11. doi: 10.1038/nrmicro2108

Morgan TH, Sturtevant AH, Muller HJ, Bridges CB (1915) The Mechanism of Mendelian Heredity. Henry Holt and Company, New York

Mounts AW, Ando T, Koopmans M, et al. (2000) Cold weather seasonality of gastroenteritis associated with Norwalk-like viruses. J Infect Dis 181:S284–7. doi: 10.1086/315586

Mullner P, Spencer SEF, Wilson DJ, et al. (2009) Assigning the source of human campylobacteriosis in New Zealand: a comparative genetic and epidemiological approach. Infect Genet Evol 9:1311–9. doi: 10.1016/j.meegid.2009.09.003

Myllykangas S, Natsoulis G, Bell JM, Ji HP (2011) Targeted sequencing library preparation by genomic DNA circularization. BMC Biotechnol 11:122. doi: 10.1186/1472-6750-11-122

Nachamkin I, Allos BM, Ho T (1998) Campylobacter species and Guillain-Barré syndrome. Clin Microbiol Rev 11:555–567.

Nachamkin I, Fischer SH, Yang X-H, et al. (1994) Immunoglobulin A antibodies directed against *Campylobacter jejuni* flagellin present in breast-milk. Epidemiol Infect 112:359–365. doi: 10.1017/S0950268800057769

Naik GA, Bhat LN, Chopade BA, Lynch JM (1994) Transfer of broad-host-range antibiotic resistance plasmids in soil microcosms. Curr Microbiol 28:209–215.

Nakano T, Lau GMG, Lau GML, et al. (2012) An updated analysis of hepatitis C virus genotypes and subtypes based on the complete coding region. Liver Int 32:339–45. doi: 10.1111/j.1478-3231.2011.02684.x

Nakano T, Lu L, He Y, et al. (2006) Population genetic history of hepatitis C virus 1b infection in China. J Gen Virol 87:73–82. doi: 10.1099/vir.0.81360-0

Nakano T, Lu L, Liu P, Pybus OG (2004) Viral gene sequences reveal the variable history of hepatitis C virus infection among countries. J Infect Dis 190:1098–108. doi: 10.1086/422606

Napoli N, Giannelli G, Parisi CV, et al. (2005) Predictive value of early virological response to treatment with different interferon-based regimens plus ribavirin in patients with chronic hepatitis C. New Microbiol 28:13–21.

Nature (2010) The sequence explosion. Nature 464:2010. doi: 10.1038/464670a

Ndjomou J, Pybus OG, Matz B (2003) Phylogenetic analysis of hepatitis C virus isolates indicates a unique pattern of endemic infection in Cameroon. J Gen Virol 84:2333–2341. doi: 10.1099/vir.0.19240-0

Nei M, Kumar S (2000) Evolutionary change of DNA sequences. Molecular Evolution and Phylogenetics. Oxford University Press, New York, pp 40–41

Neimann J, Engberg J, Mølbak K, Wegener HC (2003) A case-control study of risk factors for sporadic campylobacter infections in Denmark. Epidemiol Infect 130:353–66. doi: 10.1017/S0950268803008355

Nelson PK, Mathers BM, Cowie B, et al. (2011) Global epidemiology of hepatitis B and hepatitis C in people who inject drugs: results of systematic reviews. Lancet 378:571–83. doi: 10.1016/S0140-6736(11)61097-0

Nenonen NP, Hannoun C, Horal P, et al. (2008) Tracing of norovirus outbreak strains in mussels collected near sewage effluents. Appl Environ Microbiol 74:2544–9. doi: 10.1128/AEM.02477-07

Neu HC (1992) The crisis in antibiotic tesistance. Science 257:1064–72.

Neumann AU, Lam NP, Dahari H, et al. (1998) Hepatitis C viral dynamics in vivo and the antiviral efficacy of interferon-α therapy. Science 282:103–107. doi: 10.1126/science.282.5386.103

Newton MA, Raftery AE (1994) Approximate Bayesian inference with the weighted likelihood bootstrap. J R Stat Soc B 56:3–48.

Neyman J (1937) Outline of a theory of statistical estimation based on the classical theory of probability. Philos Trans R Soc London Ser A, Math Phys Sci 236:333–380.

Nielsen R (2002) Mapping mutations on phylogenies. Syst Biol 51:729–39. doi: 10.1080/10635150290102393

Nightingale F (1858) Notes on Matters Affecting the Health, Efficiency, and Hospital Administration of the British Army: Founded Chiefly on the Experience of the Late War. By Florence Nightingale. Presented by Request to the Secretary of State for War. Harrison and Sons, St. Martin's Lane, WC.

Nightingale F (1859) Notes on Hospitals: Being Two Papers Read Before the National Association for the Promotion of Social Science, at Liverpool, in October, 1858: With Evidence Given to the Royal Commissioners on the State of the Army in 1857, 2nd ed. John W. Parker and Son

Nightingale F (1860) Notes on Nursing: What it is, and what it is not. D. Appleton and Company, New York

Nilsson M, Hedlund K, Thorhagen M, et al. (2003) Evolution of human calicivirus RNA in vivo: accumulation of mutations in the protruding P2 domain of the capsid leads to structural changes and possibly a new phenotype. J Virol 77:13117–13124. doi: 10.1128/JVI.77.24.13117

Noel JS, Fankhauser RL, Ando T, et al. (1999) Identification of a distinct common strain of "Norwalk-like viruses" having a global distribution. J Infect Dis 179:1334–44. doi: 10.1086/314783

Nordborg M (2008) Coalescent Theory. In: Balding DJ, Bishop M, Cannings C (eds) The Handbook of Statistical Genetics, 3rd ed. John Wiley & Sons Ltd., Chichester, UK, pp 843–877

Norovirus Working Party (2012) Guidelines for the management of norovirus outbreaks in acute and community health and social care settings. http://www.hpa.org.uk/webc/hpawebfile/hpaweb_c/1317131639453. Accessed 12 Sep 2013

Nylen G, Dunstan F, Palmer SR, et al. (2002) The seasonal distribution of campylobacter infection in nine European countries and New Zealand. Epidemiol Infect 128:383–90. doi: 10.1017/S0950268802006830

O'Hagan A (1998) Eliciting expert beliefs in substantial practical applications. J R Stat Soc D-Stat 47:21–35.

Oka T, Saif LJ, Wang Q (2013) First complete genome sequence of a genogroup II genotype 18, strain QW125. Genome Announc 1:e00344–13. doi: 10.1128/genomeA.00344-13

Oliver SL, Dastjerdi AM, Wong S, et al. (2003) Molecular characterization of bovine enteric caliciviruses: a distinct third genogroup of noroviruses (Norwalk-like viruses) unlikely to be of risk to humans. J Virol 77:2789–2798. doi: 10.1128/JVI.77.4.2789

Olsen GJ, Woese CR, Overbeek R (1994) The winds of (evolutionary) change: breathing new life into microbiology. J Bacteriol 176:1–6.

Ortín J, Parra F (2006) Structure and function of RNA replication. Annu Rev Microbiol 60:305–26. doi: 10.1146/annurev.micro.60.080805.142248

Oxford University Hospitals NHS Trust (2012) Annual report 2011/12. 42:19. doi: 10.1111/epp.2606

Pallen MJ, Loman NJ, Penn CW (2010) High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. Curr Opin Microbiol 13:625–31. doi: 10.1016/j.mib.2010.08.003

Palmer S, Brown D, Morgan D (2005) Early qualitative risk assessment of the emerging zoonotic potential of animal diseases. BMJ 331:1256–60. doi: 10.1136/bmj.331.7527.1256

Palmer SR, Gully PR, White JM, et al. (1983) Water-borne outbreak of Campylobacter gastroenteritis. Lancet 321:287–290. doi: 10.1016/S0140-6736(83)91698-7

Pamilo P, Nei M (1988) Relationships between gene trees and species trees. Mol Biol Evol 5:568–583.

Pannell JR, Charlesworth B (1999) Neutral genetic diversity in a metapopulation with recurrent local extinction and recolonization. Evolution 53:664–676. doi: 10.2307/2640708

Papafragkou E, Hewitt J, Park GW, et al. (2013) Challenges of culturing human norovirus in three-dimensional organoid intestinal cell culture models. PLoS One 8:e63485. doi: 10.1371/journal.pone.0063485

Park SF (2002) The physiology of *Campylobacter* species and its relevance to their role as foodborne pathogens. Int J Food Microbiol 74:177–188. doi: 10.1016/S0168-1605(01)00678-X

Parkhill J, Wren BW, Mungall K, et al. (2000) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. Nature 403:665–8. doi: 10.1038/35001088

Parra GI, Bok K, Taylor R, et al. (2012) Immunogenicity and specificity of norovirus Consensus GII.4 virus-like particles in monovalent and bivalent vaccine formulations. Vaccine 30:3580–6. doi: 10.1016/j.vaccine.2012.03.050

Parrino TA, Schreiber DS, Trier JS, et al. (1977) Clinical immunity in acute gastroenteritis caused by Norwalk agent. N Engl J Med 297:86–89.

Parrish CR, Holmes EC, Morens DM, et al. (2008) Cross-species virus transmission and the emergence of new epidemic diseases. Microbiol Mol Biol R 72:457–70. doi: 10.1128/MMBR.00004-08

Pasteur L, Jourbert J, Chamberland C (1878) Théorie des germes et ses applications à la médecine et à la chirurgie. 1–23.

Patel MM, Hall AJ, Vinjé J, Parashar UD (2009) Noroviruses: a comprehensive review. J Clin Virol 44:1–8. doi: 10.1016/j.jcv.2008.10.009

Pattengale ND, Alipour M, Bininda-Emonds ORP, et al. (2010) How many bootstrap replicates are necessary? J Comput Biol 17:337–54. doi: 10.1089/cmb.2009.0179

Patterson C, Williams DM, Humpries CJ (1993) Congruence between molecular and morphological phylogenies. Annu Rev Ecol Syst 24:153–188.

Pawlotsky J-M (2002) Use and interpretation of virological tests for hepatitis C. Hepatology 36:S65–73. doi: 10.1053/jhep.2002.36815

Pawlotsky JM, Tsakiris L, Roudot-Thoraval F, et al. (1995) Relationship between hepatitis C virus genotypes and sources of infection in patients with chronic hepatitis C. J Infect Dis 171:1607–10. doi: 10.1093/infdis/171.6.1607

Pearson BM, Gaskin DJH, Segers RPAM, et al. (2007) The complete genome sequence of *Campylobacter jejuni* strain 81116 (NCTC11828). J Bacteriol 189:8402–3. doi: 10.1128/JB.01404-07

Pearson K (1898) Mathematical contributions to the theory of evolution. On the law of ancestral history. Proc R Soc 62:386–412.

Pebody RG, Ryan MJ, Wall PG (1997) Outbreaks of campylobacter infection: rare events for a common pathogen. Commun Dis Rep Rev 7:R33–7.

Pereira B, Milford E, Kirkman R, Levey A (1991) Transmission of hepatitis C virus by organ transplantation. N Engl J Med 325:454–460. doi: 10.1056/NEJM199108153250702

Pham DA, Leuangwutiwong P, Jittmittraphap A, et al. (2009) High prevalence of Hepatitis C virus genotype 6 in Vietnam. Asian Pac J Allergy 27:153–60.

Piddock LJ, Ricci V, Stanley K, Jones K (2000) Activity of antibiotics used in human medicine for *Campylobacter jejuni* isolated from farm animals and their environment in Lancashire, UK. J Antimicrob Chemoth 46:303–6. doi: 10.1093/jac/46.2.303

Van der Poel CL, Cuypers HTM, Reesink HW, et al. (1991) Confirmation of hepatitis C virus infection by new four-antigen recombinant immunoblot assay. Lancet 337:317–9. doi: 10.1016/0140-6736(91)90942-I

Poly F, Read T, Tribble DR, et al. (2007) Genome sequence of a clinical isolate of *Campylobacter jejuni* from Thailand. Infect Immun 75:3425–33. doi: 10.1128/IAI.00050-07

Poordad F, McCone J, Bacon BR, et al. (2011) Boceprevir for Untreated Chronic HCV Genotype 1 Infection. N Engl J Med 364:1195–1206. doi: 10.1056/NEJMoa1010494

Pope JE, Krizova A, Garg AX, et al. (2007) *Campylobacter* reactive arthritis: a systematic review. Semin Arthritis Rheu 37:48–55. doi: 10.1016/j.semarthrit.2006.12.006

Population Reference Bureau (2004) World Population Data Sheet. http://prb.org/publications/datasheets/2004/2004worldpopulationdatasheet.aspx. Accessed 30 Apr 2012

Porter JR (1973) Agostino Bassi bicentennial (1773-1973). Bacteriol Rev 37:284–8.

Posada D, Crandall KA (1998) Modeltest: testing the model of DNA substitution. Bioinformatics 14:817–818.

Pouillot R, Lachenal G, Pybus OG, et al. (2008) Variable epidemic histories of hepatitis C virus genotype 2 infection in West Africa and Cameroon. Infect Genet Evol 8:676–81. doi: 10.1016/j.meegid.2008.06.001

Powers JP, Piper DE, Li Y, et al. (2006) SAR and mode of action of novel non-nucleoside inhibitors of hepatitis C NS5b RNA polymerase. J Med Chem 49:1034–46. doi: 10.1021/jm050859x

Poynard T, Cacoub P, Ratziu V, et al. (2002) Fatigue in patients with chronic hepatitis C. J Viral Hepatitis 9:295–303. doi: 10.1046/j.1365-2893.2002.00364.x

Prati D (2006) Transmission of hepatitis C virus by blood transfusions and other medical procedures: a global review. J Hepatol 45:607–16. doi: 10.1016/j.jhep.2006.07.003

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2002) Numerical Recipes in C++. Cambridge University Press, Cambridge, UK

Prusiner SB (1982) Novel proteinaceous infectious particles cause scrapie. Science 216:136–144. doi: 10.1126/science.6801762

Prusiner SB (1991) Molecular biology of prion diseases. Science 252:1515–1522.

Pupko T, Pe'er I, Shamir R, Graur D (2000) A fast algorithm for joint reconstruction of ancestral amino acid sequences. Mol Biol Evol 17:890–6.

Pybus OG, Barnes E, Taggart R, et al. (2009) Genetic history of hepatitis C virus in East Asia. J Virol 83:1071–82. doi: 10.1128/JVI.01501-08

Pybus OG, Charleston MA, Gupta S, et al. (2001) The epidemic behavior of the hepatitis C virus. Science 292:2323–5. doi: 10.1126/science.1058321

Pybus OG, Cochrane A, Holmes EC, Simmonds P (2005) The hepatitis C virus epidemic among injecting drug users. Infect Genet Evol 5:131–9. doi: 10.1016/j.meegid.2004.08.001

Pybus OG, Drummond AJ, Nakano T, et al. (2003) The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. Mol Biol Evol 20:381–387. doi: 10.1093/molbev/msg043

Pybus OG, Rambaut A, Harvey PH (2000) An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics 155:1429–37.

Rambaut A, Bromham L (1998) Estimating divergence dates from molecular sequences. Mol Biol Evol 15:442–8.

Rambaut A, Drummond AJ (2009) Tracer v1.5. http://beast.bio.ed.ac.uk/Tracer.

Raoult D, Forterre P (2008) Redefining viruses: lessons from Mimivirus. Nat Rev Microbiol 6:315–9. doi: 10.1038/nrmicro1858

Rasko DA, Webster DR, Sahl JW, et al. (2011) Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany. N Engl J Med 365:709–717. doi: 10.1056/NEJMoa1106920

Rasmussen DA, Boni MF, Koelle K (2014) Reconciling phylodynamics with epidemiology: the case of dengue virus in southern Vietnam. Mol Biol Evol 31:258–71. doi: 10.1093/molbev/mst203

Rasmussen DA, Ratmann O, Koelle K (2011) Inference for nonlinear epidemiological models using genealogies and time series. PLoS Comput Biol 7:e1002136. doi: 10.1371/journal.pcbi.1002136

Ré VE, Culasso ACA, Mengarelli S, et al. (2011) Phylodynamics of hepatitis C virus subtype 2c in the province of Córdoba, Argentina. PLoS One 6:e19471. doi: 10.1371/journal.pone.0019471

Repp KK, Keene WE (2012) A point-source norovirus outbreak caused by exposure to fomites. J Infect Dis 205:1639–41. doi: 10.1093/infdis/jis250

Rhymer JM, Simberloff D (1996) Extinction by hybridization and introgression. Annu Rev Ecol Syst 27:83–109. doi: 10.1146/annurev.ecolsys.27.1.83

Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. Math Biosci 53:131–141. doi: 10.1016/0025-5564(81)90043-2

Robinson ER, Walker TM, Pallen MJ (2013) Genomics and outbreak investigation: from sequence to consequence. Genome Med 5:36. doi: 10.1186/gm440

Rodrigo A, Felsenstein J (1999) Coalescent approaches to HIV population genetics. In: Crandall K (ed) The Evolution of HIV. John Hopkins University Press, Baltimore, MD, pp 233–272

Rohayem J, Münch J, Rethwilm A (2005) Evidence of recombination in the norovirus capsid gene. J Virol 79:4977–4990. doi: 10.1128/JVI.79.8.4977

Rohde H, Qin J, Cui Y, et al. (2011) Open-source genomic analysis of shiga-toxin-producing *E. coli* O104:H4. N Engl J Med 365:718–24. doi: 10.1056/NEJMoa1107643

Rzhetsky A, Nei M (1995) Tests of applicability of several substitution models for DNA sequence data. Mol Biol Evol 12:131–51. doi: 10.1093/oxfordjournals.molbev.a0401

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–25.

Sam WIC, Lyons MM, Waghorn DJ (1999) Increasing rates of ciprofloxacin resistant campylobacter. J Clin Pathol 52:709. doi: 10.1080/1464727992000198551

Sanger F, Brownlee GG, Barrell BG (1965) A two-dimensional fractionation procedure for radioactive nucleotides. J Mol Biol 13:373–398. doi: 10.1016/S0022-2836(65)80104-8

Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94:441–8.

Sanger F, Donelson JE, Coulson AR, et al. (1973) Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage f1 DNA. Proc Natl Acad Sci USA 70:1209–1213.

Sanger F, Donelson JE, Coulson AR, et al. (1974) Determination of a nucleotide sequence in bacteriophage f1 DNA by primed synthesis with DNA polymerase. J Mol Biol 90:315–33.

Sanger F, Nicklen S, Coulson A (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467.

Schaeffer J, Le Saux J-C, Lora M, et al. (2013) Norovirus contamination on French marketed oysters. Int J Food Microbiol. doi: 10.1016/j.ijfoodmicro.2013.07.022

Schierup MH, Hein J (2000) Consequences of recombination on traditional phylogenetic analysis. Genetics 156:879–91.

Schleifer KH (2009) Classification of bacteria and archaea: past, present and future. Syst Appl Microbiol 32:533–42. doi: 10.1016/j.syapm.2009.09.002

Schmid D, Lederer I, Pichler A-M, et al. (2005) An outbreak of norovirus infection affecting an Austrian nursing home and a hospital. Wien Klin Wochenschr 117:802–8. doi: 10.1007/s00508-005-0473-1

Schmidt HA, von Haeseler A (2009) Phylogenetic inference using maximum likelihood methods. In: Lemey P, Salemi M, Vandamme A-M (eds) The Phylogenetic Handbook, 2nd ed. Cambridge University Press, pp 181–209

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–4. doi: 10.1093/bioinformatics/18.3.502

Schouls LM, Reulen S, Duim B, et al. (2003) Comparative genotyping of *Campylobacter jejuni* by amplified fragment length polymorphism, multilocus sequence typing, and short repeat sequencing: strain diversity, host range, and recombination. J Clin Microbiol 41:15–26. doi: 10.1128/JCM.41.1.15-26.2003

Schwartz DC, Cantor CR (1984) Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell 37:67–75. doi: 10.1016/0092-8674(84)90301-5

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464.

Sears A, Baker MG, Wilson N, et al. (2011) Marked Campylobacteriosis decline after interventions aimed at poultry, New Zealand. Emerg Infect Dis 17:1007–1015. doi: 10.3201/eid1706.101272

Sebald M, Véron M (1963) Teneur en basses de l'ADN et classification des vibrions. Ann Inst Pasteur 105:897–910.

Seng P, Drancourt M, Gouriet F, et al. (2009) Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Clin Infect Dis 49:543–51. doi: 10.1086/600885

Shapiro B, Drummond AJ, Rambaut A, et al. (2004) Rise and fall of the Beringian steppe bison. Science 306:1561–5. doi: 10.1126/science.1101074

Sharp TW, Hyams KC, Watts D, et al. (1995) Epidemiology of Norwalk virus during an outbreak of acute gastroenteritis aboard a US aircraft carrier. J Med Virol 45:61–7.

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–45. doi: 10.1038/nbt1486

Shepard CW, Finelli L, Alter MJ (2005) Global epidemiology of hepatitis C virus infection. Lancet Infect Dis 5:558–67. doi: 10.1016/S1473-3099(05)70216-4

Sheppard SK, Colles FM, McCarthy ND, et al. (2011) Niche segregation and genetic structure of *Campylobacter jejuni* populations from wild and agricultural host species. Mol Ecol 20:3484–90. doi: 10.1111/j.1365-294X.2011.05179.x

Sheppard SK, Dallas JF, Strachan NJC, et al. (2009) *Campylobacter* genotyping to determine the source of human infection. Clin Infect Dis 48:1072–8. doi: 10.1086/597402

Sheppard SK, Didelot X, Jolley KA, et al. (2013a) Progressive genome-wide introgression in agricultural *Campylobacter coli*. Mol Ecol 22:1051–64. doi: 10.1111/mec.12162

Sheppard SK, Didelot X, Meric G, et al. (2013b) Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. Proc Natl Acad Sci USA 110:11923–11927. doi: 10.1073/pnas.1305559110

Sheppard SK, Maiden MCJ, Falush D (2010) Population Genetics of *Campylobacter*. In: Robinson DA, Falush D, Feil EJ (eds) Bacterial Populations in Infectious Disease. Wiley-Blackwell, Hoboken, NJ, USA, p 184

Sheppard SK, McCarthy ND, Falush D, Maiden MCJ (2008) Convergence of *Campylobacter species*: implications for bacterial evolution. Science 320:237–9. doi: 10.1126/science.1155532

Siebenga JJ, Beersma MFC, Vennema H, et al. (2008) High prevalence of prolonged norovirus shedding and illness among hospitalized patients: a model for in vivo molecular evolution. J Infect Dis 198:994–1001. doi: 10.1086/591627

Siebenga JJ, Lemey P, Kosakovsky Pond SL, et al. (2010) Phylodynamic reconstruction reveals norovirus GII.4 epidemic expansions and their molecular determinants. PLoS Pathog 6:e1000884. doi: 10.1371/journal.ppat.1000884

Siebenga JJ, Vennema H, Renckens B, et al. (2007) Epochal evolution of GGII.4 norovirus capsid proteins from 1995 to 2006. J Virol 81:9932–41. doi: 10.1128/JVI.00674-07

Simmonds P (2004) Genetic diversity and evolution of hepatitis C virus--15 years on. J Gen Virol 85:3173–88. doi: 10.1099/vir.0.80401-0

Simmonds P, Bukh J, Combet C, et al. (2005) Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. Hepatology 42:962–73. doi: 10.1002/hep.20819

Simonsen L, Kane A, Lloyd J, et al. (1999) Unsafe injections in the developing world and transmission of bloodborne pathogens: a review. Bull World Health Organ 77:789–800.

Sjödin P, Kaj I, Krone S, et al. (2005) On the meaning and existence of an effective population size. Genetics 169:1061–70. doi: 10.1534/genetics.104.026799

Skirrow MB, Jones DM, Sutcliffe E, Benjamin J (1993) Campylobacter bacteraemia in England and Wales, 1981-91. Epidemiol Infect 110:567–73.

Slatkin M (1977) Gene flow and genetic drift in a species subject to frequent local extinctions. Theor Popul Biol 12:253–262. doi: 10.1016/0040-5809(77)90045-4

Smith GJD, Vijaykrishna D, Bahl J, et al. (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature 459:1122–5. doi: 10.1038/nature08182

Smith JM, Smith NH, O'Rourke M, Spratt BG (1993) How clonal are bacteria? Proc Natl Acad Sci USA 90:4384–8.

Smith LM, Fung S, Hunkapiller MW, et al. (1985) The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. Nucleic Acids Res 13:2399–2412. doi: 10.1093/nar/13.7.2399

Smith LM, Sanders JZ, Kaiser RJ, et al. (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321:674–679. doi: 10.1038/321674a0

Smithies O (1955) Zone electrophoresis in starch gels: group variations in the serum proteins of normal human adults. Biochem J 61:629–41.

Snow J (1855) On the Mode of Communication of Cholera, 2nd ed. Churchill, London

Sofia MJ, Bao D, Chang W, et al. (2010) Discovery of a β-d-2′-deoxy-2'-α-fluoro-2'-β-C-methyluridine nucleotide prodrug (PSI-7977) for the treatment of hepatitis C virus. J Med Chem 53:7202–18. doi: 10.1021/jm100863x

Sokal R, Michener C (1958) A statistical method for evaluating systematic relationships. Univ Kansas Sci Bull 38:1409–38.

Sperber WH, Tatini SR (1975) Interpretation of the tube coagulase test for identification of *Staphylococcus aureus*. Appl Microbiol 29:502–5.

Springthorpe VS, Sattar SA (1990) Chemical disinfection of virus-contaminated surfaces. Crit Rev Env Contr 20:169–229.

Staden R (1979) A strategy of DNA sequencing employing computer programs. Nucleic Acids Res 6:2601–2610.

Stadler T, Kouyos R, von Wyl V, et al. (2012) Estimating the basic reproductive number from viral sequence data. Mol Biol Evol 29:347–57. doi: 10.1093/molbev/msr217

Steel M (2005) Should phylogenetic models be trying to "fit an elephant"? Trends Genet 21:307–9. doi: 10.1016/j.tig.2005.04.001

Steinbrueckner B, Ruberg F, Kist M (2001) Bacterial genetic fingerprint: a reliable factor in the study of the epidemiology of human *Campylobacter* enteritis? J Appl Microbiol. doi: 10.1128/JCM.39.11.4155-4159.2001

Stonnet V, Guesdon JL (1993) *Campylobacter jejuni*: specific oligonucleotides and DNA probes for use in polymerase chain reaction-based diagnosis. FEMS Immunol Med Microbiol 7:337–44. doi: 10.1111/j.1574-695X.1993.tb00415.x

Strader DB, Wright T, Thomas DL, Seeff LB (2004) Diagnosis, management, and treatment of hepatitis C. Hepatology 39:1147–71. doi: 10.1002/hep.20119

Strimmer K, von Haeseler A (2009) Genetic distances and nucleotide substitution models. In: Lemey P, Salemi M, Vandamme A-M (eds) The Phylogenetic Handbook, 2nd ed. Cambridge University Press, Cambridge, UK, pp 111–125

Stroffolini T, Lorenzoni U, Menniti-Ippolito F, et al. (2001) Hepatitis C virus infection in spouses: sexual transmission or common exposure to the same risk factors? Am J Gastroenterol 96:3138–41. doi: 10.1111/j.1572-0241.2001.05267.x

Suárez-Díaz E, Anaya-Muñoz VH (2008) History, objectivity, and the construction of molecular phylogenies. Stud Hist Philos Biol Biomed Sci 39:451–68. doi: 10.1016/j.shpsc.2008.09.002

Suchard MA, Kitchen CMR, Sinsheimer JS, Weiss RE (2003) Hierarchical phylogenetic models for analyzing multipartite sequence data. Syst Biol 52:649–664. doi: 10.1080/10635150390238879

Sugieda M, Nagaoka H, Kakishima Y, et al. (1998) Detection of Norwalk-like virus genes in the caecum contents of pigs. Arch Virol 143:1215–1221.

Susser M (1979) Epidemiology. In: Kruskay W, Tanur J (eds) Statistics: A Volume Compiled from the International Encyclopedia of the Social Sciences, 3rd ed. Collier MacMillan, New York, pp 201–8

Swofford D (2003) PAUP*. Phylogenetic Analysis Using Parsimony (* and other methods). http://paup.csit.fsu.edu/. Accessed 31 Aug 2013

Syed F, Grunenwald H, Caruccio N (2009) Next-generation sequencing library preparation: simultaneous fragmentation and tagging using in vitro transposition. Nat Methods 6:782.

Taboada EN, van Belkum A, Yuki N, et al. (2007) Comparative genomic analysis of *Campylobacter jejuni* associated with Guillain-Barré and Miller Fisher syndromes: neuropathogenic and enteritis-associated isolates can share high levels of genomic similarity. BMC Genomics 8:359. doi: 10.1186/1471-2164-8-359

Takahata N (1989) Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics 122:957–966.

Takamiya M, Ozen A, Rasmussen M, et al. (2011) Genome Sequence of *Campylobacter jejuni* strain 327, a strain isolated from a turkey slaughterhouse. Stand Genomic Sci 4:113–22. doi: 10.4056/sigs.1313504

Talbi C, Lemey P, Suchard MA, et al. (2010) Phylodynamics and human-mediated dispersal of a zoonotic virus. PLoS Pathog 6:e1001166. doi: 10.1371/journal.ppat.1001166

Tamura K (1992) Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol Biol Evol 9:678–87.

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–26.

Tan M, Hegde RS, Jiang X (2004) The p domain of norovirus capsid protein forms dimer and binds to histo-blood group antigen receptors. J Virol 78:6233–6242. doi: 10.1128/JVI.78.12.6233

Tan M, Jiang X (2005) Norovirus and its histo-blood group antigen receptors: an answer to a historical puzzle. Trends Microbiol 13:285–93. doi: 10.1016/j.tim.2005.04.004

Tan M, Jin M, Xie H, et al. (2008) Outbreak studies of a GII-3 and a GII-4 norovirus revealed an association between HBGA phenotypes and viral infection. J Med Virol 80:1296–1301. doi: 10.1002/jmv

Tanaka Y, Hanada K, Mizokami M, et al. (2002) A comparison of the molecular clock of hepatitis C virus in the United States and Japan predicts that hepatocellular carcinoma incidence in the United States will increase over the next two decades. Proc Natl Acad Sci USA 99:15584–9. doi: 10.1073/pnas.242608099

Tanaka Y, Kurbanov F, Mano S, et al. (2006) Molecular tracing of the global hepatitis C virus epidemic predicts regional patterns of hepatocellular carcinoma mortality. Gastroenterology 130:703–14. doi: 10.1053/j.gastro.2006.01.032

Taubenberger JK, Morens DM (2006) 1918 influenza: the mother of all pandemics. Emerg Infect Dis 12:15–22. doi: 10.3201/eid1201.050979

Tavaré S (1986) Some probabilistc and statistical problems in the analysis of DNA sequences. Lectures Math Life Sci 17:57–86.

Taylor DE (1992) Genetics of *Campylobacter* and *Helicobacter*. Annu Rev Microbiol 46:35–64. doi: 10.1146/annurev.mi.46.100192.000343

Taylor DN, Echeverria P, Pitarangsi C, et al. (1988) Influence of strain characteristics and immunity on the epidemiology of *Campylobacter* infections in Thailand. J Clin Microbiol 26:863–868.

Taylor DN, Perlman DM, Echeverria PD, et al. (1993) *Campylobacter* immunity and quantitative excretion rates in Thai children. J Infect Dis 168:754–8. doi: 10.1093/infdis/168.3.754

Teunis P, Heijne JCM, Sukhrie F, et al. (2013) Infectious disease transmission as a forensic problem: who infected whom? J R Soc Interface 10:20120955. doi: 10.1098/rsif.2012.0955

Teunis PFM, Moe CL, Liu P, et al. (2008) Norwalk virus: how infectious is it? J Med Virol 80:1468–1476. doi: 10.1002/jmv

Thanbichler M, Shapiro L (2006) Chromosome organization and segregation in bacteria. J Struct Biol 156:292–303. doi: 10.1016/j.jsb.2006.05.007

Thompson CC, Chimetto L, Edwards RA, et al. (2013) Microbial genomic taxonomy. BMC Genomics 14:913. doi: 10.1186/1471-2164-14-913

Thorne H V (1966) Electrophoretic separation of polyoma virus DNA from host cell DNA. Virology 29:234–9. doi: 10.1016/0042-6822(66)90029-8

Thorne JL, Kishino H, Painter IS (1998) Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol 15:1647–57.

Tjio JH, Levan A (1956) The chromosome number of man. Hereditas 42:1–6.

Troesch M, Meunier I, Lapierre P, et al. (2006) Study of a novel hypervariable region in hepatitis C virus (HCV) E2 envelope glycoprotein. Virology 352:357–67. doi: 10.1016/j.virol.2006.05.015

Tu ET-V, Bull RA, Greening GE, et al. (2008) Epidemics of gastroenteritis during 2006 were associated with the spread of norovirus GII.4 variants 2006a and 2006b. Clin Infect Dis 46:413–20. doi: 10.1086/525259

Turcios RM, Widdowson M-A, Sulka AC, et al. (2006) Reevaluation of epidemiological criteria for identifying outbreaks of acute gastroenteritis due to norovirus: United States, 1998-2000. Clin Infect Dis 42:964–9. doi: 10.1086/500940

van Beek J, Ambert-Balay K, Botteldoorn N, et al. (2013) Indications for worldwide increased norovirus activity associated with emergence of a new variant of genotype II.4, late 2012. Euro Surveill 18:20345.

van Belkum A, van Leeuwen W, Kaufmann ME, et al. (1998) Assessment of resolution and intercenter reproducibility of results of genotyping *Staphylococcus aureus* by pulsed-field gel electrophoresis of *Sma*I macrorestriction fragments: a multicenter study. J Clin Microbiol 36:1653–9.

Vandelli C, Renzo F, Romanò L, et al. (2004) Lack of evidence of sexual transmission of hepatitis C among monogamous couples: results of a 10-year prospective follow-up study. Am J Gastroenterol 99:855–9. doi: 10.1111/j.1572-0241.2004.04150.x

Varga M, Kuntová L, Pantůček R, et al. (2012) Efficient transfer of antibiotic resistance plasmids by transduction within methicillin-resistant *Staphylococcus aureus* USA300 clone. FEMS Microbiol Lett 332:146–52. doi: 10.1111/j.1574-6968.2012.02589.x

Velge P, Cloeckaert A, Barrow P (2005) Emergence of Salmonella epidemics: The problems related to Salmonella enterica serotype Enteritidis and multiple antibiotic resistance in other major serotypes. Vet Res 36:267–288. doi: 10.1051/vetres:2005005

Véras NMC, Gray RR, Brígido LFDM, et al. (2011) High-resolution phylogenetics and phylogeography of human immunodeficiency virus type 1 subtype C epidemic in South America. J Gen Virol 92:1698–709. doi: 10.1099/vir.0.028951-0

Verbeeck J, Maes P, Lemey P, et al. (2006) Investigating the origin and spread of hepatitis C virus genotype 5a. J Virol 80:4220–4226. doi: 10.1128/JVI.80.9.4220

Vinjé J, Hamidjaja RA, Sobsey MD (2004) 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 116:109–117. doi: 10.1016/j.jviromet.2003.11.001

Vinjé J, Koopmans MPG (2000) Simultaneous detection and genotyping of "Norwalk-like viruses" by oligonucleotide array in a reverse line blot hybridization format. J Clin Microbiol 38:2595–601.

Vivancos R, Keenan A, Sopwith W, et al. (2010) Norovirus outbreak in a cruise ship sailing around the British Isles: investigation and multi-agency management of an international outbreak. J Infect 60:478–85. doi: 10.1016/j.jinf.2010.03.018

Volz E (2008) SIR dynamics in random networks with heterogeneous connectivity. J Math Biol 56:293–310. doi: 10.1007/s00285-007-0116-4

Volz EM, Kosakovsky Pond SL, Ward MJ, et al. (2009) Phylodynamics of infectious disease epidemics. Genetics 183:1421–30. doi: 10.1534/genetics.109.106021

Wagner DM, Klunk J, Harbeck M, et al. (2014) *Yersinia pestis* and the Plague of Justinian 541–543 AD: a genomic analysis. Lancet Infect Dis 3099:1–8. doi: 10.1016/S1473-3099(13)70323-2

Wakeley J (2004a) Recent trends in population genetics: more data! More math! Simple models? J Hered 95:397–405. doi: 10.1093/jhered/esh062

Wakeley J (1998) Segregating sites in Wright's island model. Theor Popul Biol 53:166–74. doi: 10.1006/tpbi.1997.1355

Wakeley J (1999) Nonequilibrium migration in human history. Genetics 153:1863–71.

Wakeley J (2001) The coalescent in an island model of population subdivision with variation among demes. Theor Popul Biol 59:133–44. doi: 10.1006/tpbi.2000.1495

Wakeley J (2004b) Metapopulation models for historical inference. Mol Ecol 13:865–875. doi: 10.1111/j.1365-294X.2004.02086.x

Wakeley J, Aliacar N (2001) Gene genealogies in a metapopulation. Genetics 159:893–905.

Wakita T, Pietschmann T, Kato T, et al. (2005) Production of infectious hepatitis C virus in tissue culture from a cloned viral genome. Nat Med 11:791–6. doi: 10.1038/nm1268

Walker AS, Eyre DW, Wyllie DH, et al. (2012) Characterisation of Clostridium difficile hospital ward-based transmission using extensive epidemiological data and molecular typing. PLoS Med 9:e1001172. doi: 10.1371/journal.pmed.1001172

Wang Q-H, Han MG, Cheetham S, et al. (2005) Porcine noroviruses related to human noroviruses. Emerg Infect Dis 11:1874–81. doi: 10.3201/eid1112.050485

Ward MJ, Lycett SJ, Kalish ML, et al. (2013) Estimating the rate of intersubtype recombination in early HIV-1 group M strains. J Virol 87:1967–73. doi: 10.1128/JVI.02478-12

Wassenaar TM, Geilhausen B, Newell DG (1998) Evidence of genomic instability in *Campylobacter jejuni* isolated from poultry. Appl Environ Microbiol 64:1816–1821.

Watson JD, Crick FHC (1953) Molecular structure of nucleic acids. Nature 171:737–738.

Wayne LG, Brenner DJ, Colwell RR, et al. (1987) Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int J Syst Bacteriol 37:463–464. doi: 10.1099/00207713-37-4-463

Webb P, Bain C (2011) Essential Epidemiology, 2nd ed. 12.

Weber DJ, Rutala WA, Miller MB, et al. (2010) Role of hospital surfaces in the transmission of emerging health care- associated pathogens: norovirus, *Clostridium difficile* , and *Acinetobacter* species. Am J Infect Control 38:S25–S33. doi: 10.1016/j.ajic.2010.04.196

Welden R (1895) Some remarks on variation in plants and animals. Proc R Soc 57:380–1.

Widdowson M-A, Cramer EH, Hadley L, et al. (2004) Outbreaks of acute gastroeneritis on cruise ships and on land: identification of a predominant circulating strain of norovirus - United States, 2002. J Infect Dis 190:27–36. doi: 10.1086/420888

Wilson DJ (2012) Insights from genomics into bacterial pathogen populations. PLoS Pathog 8:e1002874. doi: 10.1371/journal.ppat.1002874

Wilson DJ, Gabriel E, Leatherbarrow AJH, et al. (2008) Tracing the source of campylobacteriosis. PLoS Genet 4:e1000203. doi: 10.1371/journal.pgen.1000203

Wilson DJ, Gabriel E, Leatherbarrow AJH, et al. (2009) Rapid evolution and the importance of recombination to the gastroenteric pathogen *Campylobacter jejuni*. Mol Biol Evol 26:385–97. doi: 10.1093/molbev/msn264

Wilson DJ, McVean G (2006) Estimating diversifying selection and functional constraint in the presence of recombination. Genetics 172:1411–25. doi: 10.1534/genetics.105.044917

Wilson IJ, Weale ME, Balding DJ (2003) Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. J R Stat Soc A Stat 166:155–201. doi: 10.1111/1467-985X.00264

Wimalarathna HML, Richardson JF, Lawson AJ, et al. (2013) Widespread acquisition of antimicrobial resistance among *Campylobacter* isolates from UK retail poultry and evidence for clonal expansion of resistant lineages. BMC Microbiol 13:160. doi: 10.1186/1471-2180-13-160

Wingstrand A, Neimann J, Engberg J, et al. (2006) Fresh chicken as main risk factor for campylobacteriosis, Denmark. Emerg Infect Dis 12:280–284. doi: 10.3201/eid1202.050936

Wobus CE, Karst SM, Thackray LB, et al. (2004) Replication of norovirus in cell culture reveals a tropism for dendritic cells and macrophages. PLoS Biol 2:e432. doi: 10.1371/journal.pbio.0020432

Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271.

Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 74:5088–90.

Woese CR, Kandlert O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria and Eucarya. Proc Natl Acad Sci USA 87:4576–4579.

Wolf S, Williamson W, Hewitt J, et al. (2009) Molecular detection of norovirus in sheep and pigs in New Zealand farms. Vet Microbiol 133:184–9. doi: 10.1016/j.vetmic.2008.06.019

Wölk B, Sansonno D, Kräusslich H-G, et al. (2000) Subcellular localization, stability, and trans-cleavage competence of the hepatitis C virus NS3-NS4A complex expressed in tetracycline-regulated cell lines. J Virol 74:2293–2304. doi: 10.1128/JVI.74.5.2293-2304.2000

Wong THN, Dearlove BL, Hedge J, et al. (2013) Whole genome sequencing and de novo assembly identifies Sydney-like variant noroviruses and recombinants during the winter 2012/2013 outbreak in England. Virol J 10:335. doi: 10.1186/1743-422X-10-335

World Health Organisation (2004) Global database on blood safety. Report 2001–2002.

World Health Organisation (2008) Global database on blood safety. Report 2004-2005.

World Health Organisation (1999) Hepatitis C: global prevalence (update). Wkly Epidemiol Rec 74:425–427.

Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159.

Wright S (1940) Breeding structure of populations in relation to speciation. Am Nat 74:232–248. doi: 10.1086/280891

Wu Z, Sahin O, Shen Z, et al. (2013) Multi-omics approaches to deciphering a hypervirulent strain of *Campylobacter jejuni*. Genome Biol Evol 5:2217–30. doi: 10.1093/gbe/evt172

Wyatt RG, Dolin R, Blacklow NR, et al. (1974) Comparison of three agents of acute infectious nonbacterial gastroenteritis by cross-challenge in volunteers. J Infect Dis 129:709–14.

Xie W, Lewis PO, Fan Y, et al. (2011) Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst Biol 60:150–160. doi: 10.1093/sysbio/syq085

Yang X, Charlebois P, Gnerre S, et al. (2012) *De novo* assembly of highly diverse viral populations. BMC Genomics 13:475. doi: 10.1186/1471-2164-13-475

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–91. doi: 10.1093/molbev/msm088

Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J Mol Evol 39:306–314. doi: 10.1007/BF00160154

Yanofsky BYC, Carlton BC, Guest JR, et al. (1964) On the colinearity of gene structure and protein structure. Proc Natl Acad Sci USA 51:266–272.

Yao N, Reichert P, Taremi SS, et al. (1999) Molecular views of viral polyprotein processing revealed by the crystal structure of the hepatitis C virus bifunctional protease-helicase. Structure 7:1353–63. doi: 10.1016/S0969-2126(00)80025-8

Yen C, Wikswo ME, Lopman BA, et al. (2011) Impact of an emergent norovirus variant in 2009 on norovirus outbreak activity in the United States. Clin Infect Dis 53:568–71. doi: 10.1093/cid/cir478

Yoder AD, Yang Z (2000) Estimation of primate speciation dates using local molecular clocks. Mol Biol Evol 17:1081–90.

Young BC, Golubchik T, Batty EM, et al. (2012) Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. Proc Natl Acad Sci USA 109:4550–5. doi: 10.1073/pnas.1113219109

Yu M, Chuang W (2009) Treatment of chronic hepatitis C in Asia: when East meets West. J Gastroenterol Hepatol 24:336–45. doi: 10.1111/j.1440-1746.2009.05789.x

Yule GU (1924) A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Phil Trans R Soc B 213:21–87.

Zehender G, Ebranati E, Bernini F, et al. (2011) Phylogeography and epidemiological history of West Nile virus genotype 1a in Europe and the Mediterranean basin. Infect Genet Evol 11:646–53. doi: 10.1016/j.meegid.2011.02.003

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–9. doi: 10.1101/gr.074492.107

Zhao C, Ge B, De Villena J, et al. (2001) Prevalence of *Campylobacter* spp., *Escherichia coli*, and *Salmonella* serovars in retail chicken, turkey, pork, and beef from the Greater Washington, D.C., area. Appl Environ Microbiol 67:5431–5436. doi: 10.1128/AEM.67.12.5431

Zheng D-P, Ando T, Fankhauser RL, et al. (2006) Norovirus classification and proposed strain nomenclature. Virology 346:312–23. doi: 10.1016/j.virol.2005.11.015

Zhou DXM, Tang JW, Chu IMT, et al. (2006) Hepatitis C virus genotype distribution among intravenous drug user and the general population in Hong Kong. J Med Virol 78:574–581. doi: 10.1002/jmv.20578

Zhou X, Chan PKS, Tam JS, Tang JW (2011) A possible geographic origin of endemic hepatitis C virus 6a in Hong Kong: evidences for the association with Vietnamese immigration. PLoS One 6:e24889. doi: 10.1371/journal.pone.0024889

Zuckerkandl E, Pauling L (1965) Molecules as documents of evolutionary history. J Theor Biol 8:357–366.

Appendix A: Hepatitis C Sequences

Table A.1.1: Genbank accession numbers for the meta-analysis of hepatitis C.

Dataset	Accession numbers
MatBel1a	DQ363039-DQ3630083
NakBra1a	AY224831-AY224848
NakInd1a	AY224849-AY224878
NakUS1a	AY224780-AY224803
NakVie1a	AY224804-AY224830
PhaVie1a	FJ768812- FJ768831, FJ7688905
PybUK1a	AY100123-AY100193
TanUS1a	AB204625-AB204654
FuChi1b	GQ205760-GQ205855, GQ206012-GQ206036
KurMon1b	AB295119-AB295178
NakBra1b	AY224911-AY224936
NakChiA1b	AY835005-AY835039
NakChiB1b	AY834975-AY835004
NakInd1b	AY224937-AY224973
NakUS1b	AY224974-AY224990
NakVie1b	AY224879-AY224910
TanSpa1b	AB204592-AB204624
ReArg2c	JF511062- JF511136
MatBel3a	DQ363084-DQ363130
PybUK3a	AY100037-AY100041, AY100056-AY100113
TanEgy4a	AB103424-AB103457, AF271800-AF271812
HenFra5a	FN553283-FN55432
TanSoA5a	AB204662-AB204685
FuChi6a	GQ205935-GQ206010, GQ206011
PhaVie6a	FJ768777- FJ768799,FJ768906
TanHon6a	AB204686-AB204708
AkkTha6f	FJ859264- FJ859267; FJ859270- FJ859283; FJ859285, FJ859287, FJ859288, FJ859291, FJ859292, FJ859296, FJ859298, FJ859299, FJ859302, FJ859303, FJ859307, FJ859308, FJ859310, FJ859311, FJ859317, FJ859318, FJ859320, FJ859322, FJ859326, FJ859327, FJ859334
AkkTha6n	FJ859268, FJ859269, FJ859284, FJ859290, FJ859294, FJ859300, FJ859301, FJ859306, FJ859309, FJ859312- FJ859315, FJ859324, FJ859330, FJ859332

Appendix B: Mixed Model for Model Comparison

This appendix gives an outline of how the BEAST XML file should be edited to replicate the analysis from Section 3.4.3 for each of the five models: endemic, SI, SIS, SIR and model averaging approach.

B.1 Endemic Model

```
<constantSize id="coalescentEndemic" units="years">
     <populationSize>
          <parameter id="demog.popSize" value="1.0" lower="0.0"
               upper="Infinity"/>
     </populationSize>
</constantSize>
```

B.2 SI Model

```
<exponentialGrowth id="coalescentSI" units="years">
     <populationSize>
          <parameter id="demog.popSize" value="1.0" lower="0.0"
               upper="Infinity"/>
     </populationSize>
     <growthRate>
          <parameter id="demog.growthRate" value="100.0" lower="0.0"
               upper="Infinity"/>
     </growthRate>
</exponentialGrowth>
```

B.3 SIS Model

```
<logisticGrowthN0 id="coalescentSIS" units="years">
     <populationSize>
          <parameter id="demog.popSize" value="1.0" lower="0.0"
               upper="Infinity"/>
     </populationSize>
     <growthRate>
          <parameter id="demog.growthRate" value="100.0" lower="0.0"
               upper="Infinity"/>
     </growthRate>
     <t50>
```

```
          <parameter id="demog.t50" value="50.0" lower="0.0"
               upper="Infinity"/>
     </t50>
</logisticGrowthN0>
```

B.4 SIR Model

```
<epidemicSIR id="coalescentSIR" units="years" minPrevalence="0.000001">
     <populationSize>
          <parameter id="demog.popSize" value="1.0" lower="0.0"
               upper="Infinity"/>
     </populationSize>
     <growthRate>
          <parameter id="demog.growthRate" value="100.0" lower="0.0"
                    upper="Infinity"/>
     </growthRate>
     <peakTime>
          <parameter id="demog.tpeak" value="50.0" lower="0.0"
               upper="Infinity"/>
     </peakTime>
     <clearanceRate>
          <parameter id="demog.clearance" value="10.0" lower="0.0"
               upper="Infinity"/>
     </clearanceRate>
</epidemicSIR>
```

B.5 Mixed Model for Model Averaging and Model Comparison

For the analyses utilising model averaging in Section 3.5.1, the models share the same tree and evolutionary parameters, though this does not necessarily need to be the case. Careful attention must be paid to the `id` (to define a new parameter) and `idref` (to refer to an existing parameter) tags. The models are specified as below:

```
<constantSize id="coalescentEndemic" units="years">
     <populationSize>
          <parameter id="demog.popSize" value="1.0" lower="0.0"
               upper="Infinity"/>
     </populationSize>
</constantSize>
<exponentialGrowth id="coalescentSI" units="years">
     <populationSize>
          <parameter idref="demog.popSize"/>
     </populationSize>
     <growthRate>
```

```
     <parameter id="demog.growthRate" value="100.0" lower="0.0"
          upper="Infinity"/>
     </growthRate>
</exponentialGrowth>
<logisticGrowthN0 id="coalescentSIS" units="years">
     <populationSize>
          <parameter idref="demog.popSize"/>
     </populationSize>
     <growthRate>
          <parameter idref="demog.growthRate"/>
     </growthRate>
     <t50>
          <parameter id="demog.t50" value="50.0" lower="0.0"
               upper="Infinity"/>
     </t50>
</logisticGrowthN0>
<epidemicSIR id="coalescentSIR" units="years" minPrevalence="0.000001">
     <populationSize>
          <parameter idref="demog.popSize"/>
     </populationSize>
     <growthRate>
          <parameter idref="demog.growthRate"/>
     </growthRate>
     <peakTime>
          <parameter idref="demog.t50"/>
     </peakTime>
     <clearanceRate>
          <parameter id="demog.clearance" value="10.0" lower="0.0"
               upper="Infinity"/>
     </clearanceRate>
</epidemicSIR>
```

A separate coalescent likelihood calculation is required for each demographic model, for example:

```
<coalescentLikelihood id="EndemicLikelihood">
     <model>
          <constantSize idref="coalescentEndemic"/>
     </model>
     <populationTree>
          <treeModel idref="treeModel"/>
     </populationTree>
</coalescentLikelihood>
<coalescentLikelihood id="SILikelihood">
     <model>
          <constantSize idref="coalescentSI"/>
     </model>
     <populationTree>
```

```
          <treeModel idref="treeModel"/>
     </populationTree>
</coalescentLikelihood>
<coalescentLikelihood id="SISLikelihood">
     <model>
          <constantSize idref="coalescentSIS"/>
     </model>
     <populationTree>
          <treeModel idref="treeModel"/>
     </populationTree>
</coalescentLikelihood>
<coalescentLikelihood id="SIRLikelihood">
     <model>
          <constantSize idref="coalescentSIR"/>
     </model>
     <populationTree>
          <treeModel idref="treeModel"/>
     </populationTree>
</coalescentLikelihood>
```

To get the marginal likelihood across models, an `<integratedMixtureModel>` must be included in the `<prior>` section. This gives the model-averaged likelihood, with prior model probabilities being specified by the compound parameter `"modelPriors"`. This is the only place in the `<prior>` block that the coalescent likelihoods should appear. It is necessary that parameters not appearing in every model have a proper prior distribution for the analysis to be valid.

```
<prior id="prior">
...
     <integratedMixtureModel id="coalescentMixture" normalize="true">
          <coalescentLikelihood idref="EndemicLikelihood"/>
          <coalescentLikelihood idref="SILikelihood"/>
          <coalescentLikelihood idref="SISLikelihood"/>
          <coalescentLikelihood idref="SIRLikelihood"/>
          <compoundParameter id="modelPriors">
               <parameter value="0.25"/>
               <parameter value="0.25"/>
               <parameter value="0.25"/>
               <parameter value="0.25"/>
          </compoundParameter>
     </integratedMixtureModel>
</prior>
```

Finally, the coalescent likelihoods must be written to the log file so that they can be used in calculating the posterior model probabilities. Thus the log element should now resemble:

```
<log id="fileLog" logEvery="1000" fileName="mcmc.log.txt"
     overwrite="false">
...
     <coalescentLikelihood idref="EndemicLikelihood"/>
     <coalescentLikelihood idref="SILikelihood"/>
     <coalescentLikelihood idref="SISLikelihood"/>
     <coalescentLikelihood idref="SIRLikelihood"/>
</log>
```

Appendix C: GII.4 Norovirus Sequences from Genbank

An asterisk (*) indicates sequences used as the reference genome for major strain type of GII.4 norovirus.

Accession	Year	Sequence Length	Description
AB220921.1	2005	7535	Norovirus Hu/Chiba/04-1050/2005/JP
AB220922.1	2005	7559	Norovirus Hu/Sakai/04-179/2005/JP
AB447427.1	2006	7533	Norovirus Hu/GII-4/Hokkaido1/2006/JP
AB447428.1	2006	7511	Norovirus Hu/GII-4/Hokkaido2/2006/JP
AB447429.1	2006	7511	Norovirus Hu/GII-4/Hokkaido3/2006/JP
AB447430.1	2006	7511	Norovirus Hu/GII-4/Hokkaido4/2006/JP
AB447431.1	2006	7511	Norovirus Hu/GII-4/Hokkaido5/2006/JP
AB447434.1	2006	7511	Norovirus Hu/GII-4/Aomori4/2006/JP
AB447435.1	2006	7511	Norovirus Hu/GII-4/Aomori5/2006/JP
AB447436.1	2006	7511	Norovirus Hu/GII-4/Akita1/2006/JP
AB447437.1	2006	7511	Norovirus Hu/GII-4/Akita2/2006/JP
AB447438.1	2006	7511	Norovirus Hu/GII-4/Akita4/2006/JP
AB447439.1	2006	7511	Norovirus Hu/GII-4/Akita5/2006/JP
AB447440.1	2006	7511	Norovirus Hu/GII-4/Miyagi2/2006/JP
AB447441.1	2006	7511	Norovirus Hu/GII-4/Miyagi4/2006/JP
AB447442.1	2006	7511	Norovirus Hu/GII-4/Miyagi5/2006/JP
AB447443.1	2006	7511	Norovirus Hu/GII-4/Toyama1/2006/JP
AB447444.1	2006	7511	Norovirus Hu/GII-4/Toyama4/2006/JP
AB447445.1	2006	7511	Norovirus Hu/GII-4/Toyama5/2006/JP
AB447446.1	2006	7511	Norovirus Hu/GII-4/Aichi3/2006/JP
AB447447.1	2006	7511	Norovirus Hu/GII-4/Aichi4/2006/JP
AB447448.1	2006	7511	Norovirus Hu/GII-4/Sakai2/2006/JP
AB447449.1	2006	7511	Norovirus Hu/GII-4/Sakai3/2006/JP
AB447450.1	2006	7511	Norovirus Hu/GII-4/Sakai4/2006/JP
AB447451.1	2006	7511	Norovirus Hu/GII-4/Hiroshima1/2006/JP
AB447452.1	2006	7511	Norovirus Hu/GII-4/Hiroshima2/2006/JP
AB447453.1	2006	7511	Norovirus Hu/GII-4/Ehime1/2006/JP
AB447454.1	2006	7511	Norovirus Hu/GII-4/Ehime2/2006/JP
AB447455.1	2006	7511	Norovirus Hu/GII-4/Ehime5/2006/JP
AB447456.1	2006	7511	Norovirus Hu/GII-4/Saga1/2006/JP
AB447457.1	2006	7511	Norovirus Hu/GII-4/Saga4/2006/JP
AB447458.1	2006	7511	Norovirus Hu/GII-4/Saga5/2006/JP
AB447459.1	2006	7511	Norovirus Hu/GII-4/Kumamoto1/2006/JP
AB447460.1	2006	7511	Norovirus Hu/GII-4/Kumamoto2/2006/JP
AB447461.1	2006	7511	Norovirus Hu/GII-4/Kumamoto3/2006/JP

AB447462.1	2006	7511	Norovirus Hu/GII-4/Kumamoto4/2006/JP
AB447463.1	2006	7511	Norovirus Hu/GII-4/Kumamoto5/2006/JP
AB541201.1	2007	7511	Norovirus Hu/GII-4/Aichi1/2007/JP
AB541202.1	2008	7509	Norovirus Hu/GII-4/Aichi1/2008/JP
AB541203.1	2007	7509	Norovirus Hu/GII-4/Aichi2/2007/JP
AB541204.1	2008	7509	Norovirus Hu/GII-4/Aichi2/2008/JP
AB541205.1	2007	7509	Norovirus Hu/GII-4/Aichi3/2007/JP
AB541206.1	2008	7509	Norovirus Hu/GII-4/Aichi3/2008/JP
AB541207.1	2007	7509	Norovirus Hu/GII-4/Aichi4/2007/JP
AB541208.1	2008	7509	Norovirus Hu/GII-4/Aichi4/2008/JP
AB541209.1	2007	7509	Norovirus Hu/GII-4/Aichi5/2007/JP
AB541210.1	2008	7509	Norovirus Hu/GII-4/Aichi5/2008/JP
AB541211.1	2007	7509	Norovirus Hu/GII-4/Akita1/2007/JP
AB541212.1	2008	7509	Norovirus Hu/GII-4/Akita1/2008/JP
AB541213.1	2008	7404	Norovirus Hu/GII-4/Akita2/2008/JP
AB541214.1	2007	7509	Norovirus Hu/GII-4/Akita3/2007/JP
AB541215.1	2008	7509	Norovirus Hu/GII-4/Akita3/2008/JP
AB541216.1	2008	7509	Norovirus Hu/GII-4/Akita4/2008/JP
AB541217.1	2007	7509	Norovirus Hu/GII-4/Akita5/2007/JP
AB541218.1	2007	7509	Norovirus Hu/GII-4/Aomori1/2007/JP
AB541219.1	2008	7509	Norovirus Hu/GII-4/Aomori1/2008/JP
AB541220.1	2007	7509	Norovirus Hu/GII-4/Aomori2/2007/JP
AB541221.1	2008	7509	Norovirus Hu/GII-4/Aomori2/2008/JP
AB541222.1	2007	7404	Norovirus Hu/GII-4/Aomori3/2007/JP
AB541223.1	2008	7509	Norovirus Hu/GII-4/Aomori3/2008/JP
AB541224.1	2007	7509	Norovirus Hu/GII-4/Aomori4/2007/JP
AB541225.1	2008	7509	Norovirus Hu/GII-4/Aomori4/2008/JP
AB541226.1	2007	7509	Norovirus Hu/GII-4/Aomori5/2007/JP
AB541227.1	2008	7509	Norovirus Hu/GII-4/Aomori5/2008/JP
AB541228.1	2007	7509	Norovirus Hu/GII-4/Chiba1/2007/JP
AB541229.1	2008	7509	Norovirus Hu/GII-4/Chiba1/2008/JP
AB541230.1	2007	7509	Norovirus Hu/GII-4/Chiba2/2007/JP
AB541231.1	2008	7509	Norovirus Hu/GII-4/Chiba2/2008/JP
AB541232.1	2007	7509	Norovirus Hu/GII-4/Chiba4/2007/JP
AB541233.1	2008	7509	Norovirus Hu/GII-4/Chiba4/2008/JP
AB541234.1	2007	7497	Norovirus Hu/GII-4/Chiba5/2007/JP
AB541235.1	2008	7509	Norovirus Hu/GII-4/Chiba5/2008/JP
AB541236.1	2007	7509	Norovirus Hu/GII-4/Ehime1/2007/JP
AB541237.1	2008	7509	Norovirus Hu/GII-4/Ehime1/2008/JP
AB541238.1	2007	7509	Norovirus Hu/GII-4/Ehime2/2007/JP
AB541239.1	2007	7509	Norovirus Hu/GII-4/Ehime3/2007/JP
AB541240.1	2008	7509	Norovirus Hu/GII-4/Ehime3/2008/JP
AB541241.1	2007	7509	Norovirus Hu/GII-4/Ehime4/2007/JP
AB541242.1	2008	7509	Norovirus Hu/GII-4/Ehime4/2008/JP
AB541243.1	2008	7509	Norovirus Hu/GII-4/Ehime5/2008/JP
AB541244.1	2008	7509	Norovirus Hu/GII-4/Fukui1/2008/JP

AB541245.1	2007	7509	Norovirus Hu/GII-4/Fukui2/2007/JP
AB541246.1	2008	7509	Norovirus Hu/GII-4/Fukui2/2008/JP
AB541247.1	2007	7509	Norovirus Hu/GII-4/Fukui4/2007/JP
AB541248.1	2008	7509	Norovirus Hu/GII-4/Fukui4/2008/JP
AB541249.1	2007	7509	Norovirus Hu/GII-4/Fukui5/2007/JP
AB541250.1	2008	7509	Norovirus Hu/GII-4/Fukui5/2008/JP
AB541251.1	2007	7509	Norovirus Hu/GII-4/Hiroshima1/2007/JP
AB541252.1	2008	7509	Norovirus Hu/GII-4/Hiroshima1/2008/JP
AB541253.1	2007	7509	Norovirus Hu/GII-4/Hiroshima2/2007/JP
AB541254.1	2008	7509	Norovirus Hu/GII-4/Hiroshima2/2008/JP
AB541255.1	2007	7509	Norovirus Hu/GII-4/Hiroshima3/2007/JP
AB541256.1	2008	7509	Norovirus Hu/GII-4/Hiroshima3/2008/JP
AB541257.1	2007	7509	Norovirus Hu/GII-4/Hiroshima4/2007/JP
AB541258.1	2008	7509	Norovirus Hu/GII-4/Hiroshima4/2008/JP
AB541259.1	2008	7509	Norovirus Hu/GII-4/Hiroshima5/2008/JP
AB541260.1	2007	7509	Norovirus Hu/GII-4/Hokkaido1/2007/JP
AB541261.1	2008	7509	Norovirus Hu/GII-4/Hokkaido1/2008/JP
AB541262.1	2007	7509	Norovirus Hu/GII-4/Hokkaido2/2007/JP
AB541263.1	2008	7509	Norovirus Hu/GII-4/Hokkaido2/2008/JP
AB541264.1	2008	7509	Norovirus Hu/GII-4/Hokkaido3/2008/JP
AB541265.1	2007	7509	Norovirus Hu/GII-4/Hokkaido4/2007/JP
AB541266.1	2008	7509	Norovirus Hu/GII-4/Hokkaido4/2008/JP
AB541267.1	2007	7509	Norovirus Hu/GII-4/Hokkaido5/2007/JP
AB541268.1	2008	7509	Norovirus Hu/GII-4/Hokkaido5/2008/JP
AB541269.1	2008	7509	Norovirus Hu/GII-4/Iwate1/2008/JP
AB541270.1	2007	7509	Norovirus Hu/GII-4/Iwate2/2007/JP
AB541271.1	2007	7509	Norovirus Hu/GII-4/Iwate3/2007/JP
AB541272.1	2008	7509	Norovirus Hu/GII-4/Iwate3/2008/JP
AB541273.1	2007	7509	Norovirus Hu/GII-4/Iwate4/2007/JP
AB541274.1	2008	7509	Norovirus Hu/GII-4/Iwate4/2008/JP
AB541275.1	2007	7509	Norovirus Hu/GII-4/Iwate5/2007/JP
AB541276.1	2008	7506	Norovirus Hu/GII-4/Iwate5/2008/JP
AB541277.1	2007	7509	Norovirus Hu/GII-4/Kumamoto1/2007/JP
AB541278.1	2007	7509	Norovirus Hu/GII-4/Kumamoto2/2007/JP
AB541279.1	2007	7509	Norovirus Hu/GII-4/Kumamoto3/2007/JP
AB541280.1	2007	7509	Norovirus Hu/GII-4/Kumamoto4/2007/JP
AB541281.1	2007	7509	Norovirus Hu/GII-4/Miyagi1/2007/JP
AB541282.1	2007	7509	Norovirus Hu/GII-4/Miyagi2/2007/JP
AB541283.1	2007	7509	Norovirus Hu/GII-4/Miyagi3/2007/JP
AB541284.1	2008	7509	Norovirus Hu/GII-4/Miyagi5/2008/JP
AB541285.1	2008	7509	Norovirus Hu/GII-4/Miyazaki10/2008/JP
AB541286.1	2008	7509	Norovirus Hu/GII-4/Miyazaki12/2008/JP
AB541287.1	2008	7509	Norovirus Hu/GII-4/Miyazaki13/2008/JP
AB541288.1	2007	7509	Norovirus Hu/GII-4/Miyazaki1/2007/JP
AB541289.1	2008	7509	Norovirus Hu/GII-4/Miyazaki1/2008/JP
AB541290.1	2007	7509	Norovirus Hu/GII-4/Miyazaki2/2007/JP

AB541291.1	2008	7509	Norovirus Hu/GII-4/Miyazaki2/2008/JP
AB541292.1	2007	7509	Norovirus Hu/GII-4/Miyazaki3/2007/JP
AB541293.1	2008	7509	Norovirus Hu/GII-4/Miyazaki3/2008/JP
AB541294.1	2007	7509	Norovirus Hu/GII-4/Miyazaki4/2007/JP
AB541295.1	2008	7509	Norovirus Hu/GII-4/Miyazaki4/2008/JP
AB541296.1	2007	7509	Norovirus Hu/GII-4/Miyazaki5/2007/JP
AB541297.1	2008	7509	Norovirus Hu/GII-4/Miyazaki6/2008/JP
AB541298.1	2008	7509	Norovirus Hu/GII-4/Miyazaki7/2008/JP
AB541299.1	2008	7509	Norovirus Hu/GII-4/Miyazaki8/2008/JP
AB541300.1	2008	7509	Norovirus Hu/GII-4/Miyazaki9/2008/JP
AB541301.1	2007	7263	Norovirus Hu/GII-4/Nagano1/2007/JP
AB541302.1	2008	7509	Norovirus Hu/GII-4/Nagano1/2008/JP
AB541303.1	2007	7509	Norovirus Hu/GII-4/Nagano2/2007/JP
AB541304.1	2008	7509	Norovirus Hu/GII-4/Nagano2/2008/JP
AB541305.1	2007	7509	Norovirus Hu/GII-4/Nagano3/2007/JP
AB541306.1	2008	7512	Norovirus Hu/GII-4/Nagano3/2008/JP
AB541307.1	2008	7509	Norovirus Hu/GII-4/Nagano4/2008/JP
AB541308.1	2007	7509	Norovirus Hu/GII-4/Nagano5/2007/JP
AB541309.1	2007	7509	Norovirus Hu/GII-4/Niigata1/2007/JP
AB541310.1	2008	7509	Norovirus Hu/GII-4/Niigata1/2008/JP
AB541311.1	2007	7509	Norovirus Hu/GII-4/Niigata2/2007/JP
AB541312.1	2008	7509	Norovirus Hu/GII-4/Niigata2/2008/JP
AB541313.1	2007	7509	Norovirus Hu/GII-4/Niigata3/2007/JP
AB541314.1	2008	7509	Norovirus Hu/GII-4/Niigata3/2008/JP
AB541315.1	2007	7509	Norovirus Hu/GII-4/Niigata4/2007/JP
AB541316.1	2008	7509	Norovirus Hu/GII-4/Niigata4/2008/JP
AB541317.1	2007	7509	Norovirus Hu/GII-4/Niigata5/2007/JP
AB541318.1	2008	7509	Norovirus Hu/GII-4/Niigata5/2008/JP
AB541319.1*	2007	7509	Norovirus Hu/GII-4/Osaka1/2007/JP
AB541320.1	2008	7509	Norovirus Hu/GII-4/Osaka1/2008/JP
AB541321.1	2007	7509	Norovirus Hu/GII-4/Osaka2/2007/JP
AB541322.1	2008	7509	Norovirus Hu/GII-4/Osaka2/2008/JP
AB541323.1	2007	7509	Norovirus Hu/GII-4/Osaka3/2007/JP
AB541324.1	2008	7509	Norovirus Hu/GII-4/Osaka3/2008/JP
AB541325.1	2007	7509	Norovirus Hu/GII-4/Osaka4/2007/JP
AB541326.1	2008	7509	Norovirus Hu/GII-4/Osaka4/2008/JP
AB541327.1	2007	7509	Norovirus Hu/GII-4/Osaka5/2007/JP
AB541328.1	2008	7509	Norovirus Hu/GII-4/Osaka5/2008/JP
AB541329.1	2008	7509	Norovirus Hu/GII-4/Osaka6/2008/JP
AB541330.1	2007	7509	Norovirus Hu/GII-4/Saga1/2007/JP
AB541331.1	2008	7509	Norovirus Hu/GII-4/Saga1/2008/JP
AB541332.1	2007	7509	Norovirus Hu/GII-4/Saga2/2007/JP
AB541333.1	2008	7509	Norovirus Hu/GII-4/Saga2/2008/JP
AB541334.1	2008	7509	Norovirus Hu/GII-4/Saga3/2008/JP
AB541335.1	2007	7509	Norovirus Hu/GII-4/Saga4/2007/JP
AB541336.1	2008	7509	Norovirus Hu/GII-4/Saga4/2008/JP

AB541337.1	2007	7509	Norovirus Hu/GII-4/Saga5/2007/JP
AB541338.1	2008	7509	Norovirus Hu/GII-4/Saga5/2008/JP
AB541339.1	2007	7509	Norovirus Hu/GII-4/Sakai1/2007/JP
AB541340.1	2008	7509	Norovirus Hu/GII-4/Sakai1/2008/JP
AB541341.1	2007	7509	Norovirus Hu/GII-4/Sakai2/2007/JP
AB541342.1	2007	7509	Norovirus Hu/GII-4/Sakai3/2007/JP
AB541343.1	2008	7509	Norovirus Hu/GII-4/Sakai3/2008/JP
AB541344.1	2007	7509	Norovirus Hu/GII-4/Sakai4/2007/JP
AB541345.1	2008	7509	Norovirus Hu/GII-4/Sakai4/2008/JP
AB541346.1	2007	7509	Norovirus Hu/GII-4/Shimane1/2007/JP
AB541347.1	2007	7509	Norovirus Hu/GII-4/Shimane2/2007/JP
AB541348.1	2008	7509	Norovirus Hu/GII-4/Shimane2/2008/JP
AB541349.1	2007	7401	Norovirus Hu/GII-4/Shimane3/2007/JP
AB541350.1	2008	7509	Norovirus Hu/GII-4/Shimane3/2008/JP
AB541351.1	2007	7509	Norovirus Hu/GII-4/Shimane4/2007/JP
AB541352.1	2007	7509	Norovirus Hu/GII-4/Shimane5/2007/JP
AB541353.1	2008	7509	Norovirus Hu/GII-4/Shimane5/2008/JP
AB541354.1	2007	7509	Norovirus Hu/GII-4/Toyama1/2007/JP
AB541355.1	2007	7509	Norovirus Hu/GII-4/Toyama2/2007/JP
AB541356.1	2008	7509	Norovirus Hu/GII-4/Toyama2/2008/JP
AB541357.1	2007	7509	Norovirus Hu/GII-4/Toyama3/2007/JP
AB541358.1	2008	7512	Norovirus Hu/GII-4/Toyama3/2008/JP
AB541359.1	2007	7509	Norovirus Hu/GII-4/Toyama4/2007/JP
AB541360.1	2008	7509	Norovirus Hu/GII-4/Toyama4/2008/JP
AB541361.1	2007	7509	Norovirus Hu/GII-4/Toyama5/2007/JP
AB541362.1	2008	7509	Norovirus Hu/GII-4/Toyama5/2008/JP
AB543808.1	2010	7509	Norovirus Hu/GII-4/FUMI/2010/JP
AF145896.1*	1994	7509	Camberwell virus
AY032605.1	1987	7555	Human calicivirus Hu/NLV/GII/MD145-12/1987/US
AY502023.1*	2002	7556	Norovirus Hu/NoV/Farmington Hills/2002/USA
AY485642.1	2002	7559	Human calicivirus NLV/GII/Langen1061/2002/DE
AY581254.1	2003	7558	Human calicivirus Hu/NLV/Oxford/B5S22/2003/UK
AY587983.1	2002	7558	Norovirus Hu/NLV/Oxford/B4S2/2002/UK
AY587984.1	2002	7558	Norovirus Hu/NLV/Oxford/B4S5/2002/UK
AY587985.1	2002	7558	Norovirus Hu/NLV/Oxford/B4S6/2002/UK
AY587986.1	2002	7558	Norovirus Hu/NLV/Oxford/B4S4/2002/UK
AY587987.1	2002	7558	Norovirus Hu/NLV/Oxford/B4S7/2002/UK
AY587988.1	2002	7558	Norovirus Hu/NLV/Oxford/B4S1/2002/UK
AY587989.1	2002	7558	Norovirus Hu/NLV/Oxford/B2S16/2002/UK
AY741811.1*	1997	7558	Norovirus Hu/NLV/Dresden174/pUS-NorII/1997/GE
DQ078814.2*	2004	7555	Norovirus Hu/GII.4/Hunter504D/04O/AU
DQ369797.1*	2003	7559	Norovirus Hu/Guangzhou/NVgz01/CHN
DQ658413.1	2004	7558	Norovirus Hu/GII.4/MD-2004/2004/US
EF187497.2*	2006	7558	Norovirus Hu/GII.4/Kenepuru/NZ327/2006/NZL
EF684915.2*	2006	7559	Norovirus Hu/GII.4/Shellharbour/NSW696T/2006/AUS
EU310927.1*	2002	7560	Norovirus Hu/Houston/TCH186/2002/US

EU921344.2	2006	7559	Norovirus Hu/Pune/PC15/2006/India
EU921388.2	2007	7559	Norovirus Hu/Pune/PC51/2007/India
FJ514242.1	2008	7559	Norovirus Hu/GII-4/CUK-3/2008/KR
FJ537134.1	1974	7559	Norovirus Hu/GII.4/CHDC5191/1974/US
FJ537135.1	1974	7580	Norovirus Hu/GII.4/CHDC2094/1974/US
FJ537136.1*	1988	7576	Norovirus Hu/GII.4/CHDC3967/1988/US
FJ537137.1	1987	7580	Norovirus Hu/GII.4/CHDC4108/1987/US
FJ537138.1	1977	7580	Norovirus Hu/GII.4/CHDC4871/1977/US
GQ845024.2	2007	7580	Norovirus Hu/GII.4/Rathmines/NSW287R/2007/AUS
GQ845366.2	2008	7562	Norovirus Hu/GII.4/Westmead/NSW3639/2008/AUS
GQ845367.2*	2008	7560	Norovirus Hu/GII.4/Orange/NSW001P/2008/AUS
GQ845368.2*	2007	7559	Norovirus Hu/GII.4/Sutherland/NSW505G/2007/AUS
GQ845369.3	2008	7559	Norovirus Hu/GII.4/Armidale/NSW390I/2008/AUS
GU325839.2	2009	7560	Norovirus Hu/GII.4/HS194/2009/US
GU445325.2	2009	7560	Norovirus Hu/GII.4/New Orleans1805/2009/USA
GU991353.1	2008	7559	Norovirus Hu/GII/Shanghai/SH2/2008/CHN
GU991354.1	2009	7555	Norovirus Hu/Shanghai/SH5/2009/CHN
HM748971.2	2009	7511	Norovirus Hu/GII.4/Beecroft/NSW305P/2009/AUS
HM748972.2	2009	7560	Norovirus Hu/GII.4/Teralba/NSW881Z/2009/AUS
HM748973.2	2009	7559	Norovirus Hu/GII.4/Turramurra/NSW892U/2009/AUS
HQ009513.1*	2008	7560	Norovirus Hu/GII.4/JB-15/KOR/2008
JN400599.1	2006	7558	Norovirus Hu/GII-4/CGMH01/2006/TW
JN400600.1	2006	7509	Norovirus Hu/GII-4/CGMH02/2006/TW
JN400601.1	2006	7509	Norovirus Hu/GII-4/CGMH03/2006/TW
JN400602.1	2006	7509	Norovirus Hu/GII-4/CGMH04/2006/TW
JN400603.1	2006	7509	Norovirus Hu/GII-4/CGMH05/2006/TW
JN400604.1	2006	7509	Norovirus Hu/GII-4/CGMH06/2006/TW
JN400605.1	2006	7509	Norovirus Hu/GII-4/CGMH07/2006/TW
JN400606.1	2006	7509	Norovirus Hu/GII-4/CGMH08/2006/TW
JN400607.1	2006	7509	Norovirus Hu/GII-4/CGMH09/2006/TW
JN400608.1	2006	7509	Norovirus Hu/GII-4/CGMH10/2006/TW
JN400609.1	2006	7509	Norovirus Hu/GII-4/CGMH11/2006/TW
JN400610.1	2007	7509	Norovirus Hu/GII-4/CGMH12/2007/TW
JN400611.1	2007	7509	Norovirus Hu/GII-4/CGMH13/2007/TW
JN400612.1	2007	7509	Norovirus Hu/GII-4/CGMH14/2007/TW
JN400613.1	2007	7509	Norovirus Hu/GII-4/CGMH15/2007/TW
JN400614.1	2007	7509	Norovirus Hu/GII-4/CGMH16/2007/TW
JN400615.1	2007	7509	Norovirus Hu/GII-4/CGMH17/2007/TW
JN400616.1	2008	7509	Norovirus Hu/GII-4/CGMH18/2008/TW
JN400617.1	2009	7509	Norovirus Hu/GII-4/CGMH19/2009/TW
JN400618.1	2009	7509	Norovirus Hu/GII-4/CGMH20/2009/TW
JN400619.1	2010	7476	Norovirus Hu/GII-4/CGMH21/2010/TW
JN400620.1	2010	7500	Norovirus Hu/GII-4/CGMH22/2010/TW
JN400621.1	2010	7509	Norovirus Hu/GII-4/CGMH23/2010/TW
JN400622.1	2010	7509	Norovirus Hu/GII-4/CGMH24/2010/TW
JN400623.1	2010	7509	Norovirus Hu/GII-4/CGMH25/2010/TW

JN400624.1	2010	7509	Norovirus Hu/GII-4/CGMH26/2010/TW
JN400625.1	2010	7509	Norovirus Hu/GII-4/CGMH27/2010/TW
JN400626.1	2010	7509	Norovirus Hu/GII-4/CGMH28/2010/TW
JN595867.1	2010	7509	Norovirus Hu/GII.4/New Orleans/2010/USA
JQ613552.2	2010	7559	Norovirus Hu/GII.4/NSW123B/2010/AU
JQ613570.1	2009	7559	Norovirus Hu/GII.4/Rockdale/NSW006D/2009/AU
JQ613571.1	2010	7559	Norovirus Hu/GII.4/Miranda/NSW817L/2010/AU
JQ613572.1	2010	7559	Norovirus Hu/GII.4/StVincents/NSW217I/2010/AU
JQ613573.1	2010	7560	Norovirus Hu/GII.4/Helensburgh/NSW295E/2010/AU
JQ622197.1	2007	7559	Norovirus Hu/GII-4/CBNU2/2007/KR
JQ798158.1	2004	7583	Norovirus Hu/GII.4/5M/USA/2004
JQ911595.1	2009	7558	Norovirus Hu/GII/10002/2009/VNM
JQ911596.1	2009	7511	Norovirus Hu/GII/10003/2009/VNM
JQ911597.1	2009	7511	Norovirus Hu/GII/10012/2009/VNM
JQ911598.1	2009	7510	Norovirus Hu/GII/10037/2009/VNM, complete genome.
JX023286.1	1974	7511	Norovirus Hu/GII.4/CHDC5191/1974/USA
JX126912.1	2012	7549	Norovirus Hu/GII.4/Ohio/7I/2012/USA
JX126913.1	2012	7558	Norovirus Hu/GII.4/Ohio/7G/2012/USA
JX439815.1	2010	7559	Norovirus Hu/GII/Seoul1055/KOR/2010
JX439816.1	2010	7538	Norovirus Hu/GII/Seoul1072/KOR/2010
JX439817.1	2010	7537	Norovirus Hu/GII/Seoul1282/KOR/2010
JX439818.1	2010	7538	Norovirus Hu/GII/Seoul1367/KOR/2010
JX439819.1	2011	7537	Norovirus Hu/GII/Seoul1488/KOR/2011
JX448566.1	2010	7538	Norovirus Hu/GII.4/Seoul/1071/2010/KOR
JX459900.1	2011	7538	Norovirus Hu/GII.4/Randwick/NSW882J/2011/AU
JX459901.1	2011	7558	Norovirus Hu/GII.4/Caringbah/NSW409G/2011/AU
JX459902.1	2012	7559	Norovirus Hu/GII.4/Berowra/NSW767L/2012/AU
JX459903.1	2011	7559	Norovirus Hu/GII.4/Jannali/NSW774M/2011/AU
JX459904.1	2011	7559	Norovirus Hu/GII.4/Doonside/NSW536I/2011/AU
JX459905.1	2011	7559	Norovirus Hu/GII.4/Randwick/NSW938K/2011/AU
JX459906.1	2011	7560	Norovirus Hu/GII.4/Miranda/NSW850K/2011/AU
JX459907.1	2012	7560	Norovirus Hu/GII.4/Woonona/NSW3309/2012/AU
JX459908.1*	2012	7560	Norovirus Hu/GII.4/Sydney/NSW0514/2012/AU
JX989073.1	2010	7564	Norovirus Hu/GII.4/GZ2010-L26/Guangzhou/CHN/2010
JX989074.1	2011	7559	Norovirus Hu/GII.4/GZ2010-L87/Guangzhou/CHN/2011
KC013592.1	2004	7559	Norovirus Hu/GII.4/HS191/2004/USA
KC175323.1*	2012	7556	Norovirus Hu/GII.4/Hong Kong/CUHK3630/2012/CHN
KC175342.1	2009	7559	Norovirus Hu/Norwalk/10034/2009/VNM
KC175343.1	2009	7511	Norovirus Hu/Norwalk/10051/2009/VNM
KC175344.1	2009	7511	Norovirus Hu/Norwalk/10054/2009/VNM
KC175345.1	2009	7511	Norovirus Hu/Norwalk/10062/2009/VNM
KC175346.1	2009	7511	Norovirus Hu/Norwalk/10074/2009/VNM
KC175347.1	2009	7511	Norovirus Hu/Norwalk/10075/2009/VNM
KC175348.1	2009	7511	Norovirus Hu/Norwalk/10078/2009/VNM
KC175349.1	2009	7511	Norovirus Hu/Norwalk/10079/2009/VNM
KC175350.1	2009	7511	Norovirus Hu/Norwalk/10110/2009/VNM

KC175351.1	2009	7511	Norovirus Hu/Norwalk/10114/2009/VNM
KC175352.1	2009	7511	Norovirus Hu/Norwalk/10116/2009/VNM
KC175353.1	2009	7511	Norovirus Hu/Norwalk/10129/2009/VNM
KC175354.1	2009	7511	Norovirus Hu/Norwalk/10136/2009/VNM
KC175355.1	2009	7511	Norovirus Hu/Norwalk/10137/2009/VNM
KC175356.1	2009	7511	Norovirus Hu/Norwalk/10145/2009/VNM
KC175357.1	2009	7511	Norovirus Hu/Norwalk/10148/2009/VNM
KC175358.1	2009	7511	Norovirus Hu/Norwalk/10158/2009/VNM
KC175359.1	2009	7511	Norovirus Hu/Norwalk/10160/2009/VNM
KC175360.1	2009	7511	Norovirus Hu/Norwalk/10162/2009/VNM
KC175361.1	2009	7511	Norovirus Hu/Norwalk/10163/2009/VNM
KC175362.1	2009	7511	Norovirus Hu/Norwalk/10169/2009/VNM
KC175363.1	2009	7511	Norovirus Hu/Norwalk/10173/2009/VNM
KC175364.1	2009	7511	Norovirus Hu/Norwalk/10176/2009/VNM
KC175365.1	2009	7511	Norovirus Hu/Norwalk/10177/2009/VNM
KC175366.1	2009	7511	Norovirus Hu/Norwalk/10179/2009/VNM
KC175367.1	2009	7511	Norovirus Hu/Norwalk/10182/2009/VNM
KC175368.1	2009	7511	Norovirus Hu/Norwalk/10183/2009/VNM
KC175369.1	2009	7511	Norovirus Hu/Norwalk/10194/2009/VNM
KC175370.1	2009	7511	Norovirus Hu/Norwalk/10199/2009/VNM
KC175371.1	2009	7511	Norovirus Hu/Norwalk/10203/2009/VNM
KC175372.1	2009	7511	Norovirus Hu/Norwalk/10204/2009/VNM
KC175373.1	2009	7511	Norovirus Hu/Norwalk/10222/2009/VNM
KC175374.1	2009	7511	Norovirus Hu/Norwalk/10223/2009/VNM
KC175375.1	2009	7511	Norovirus Hu/Norwalk/10235/2009/VNM
KC175376.1	2009	7511	Norovirus Hu/Norwalk/10236/2009/VNM
KC175377.1	2009	7511	Norovirus Hu/Norwalk/10238/2009/VNM
KC175378.1	2009	7511	Norovirus Hu/Norwalk/10247/2009/VNM
KC175379.1	2009	7511	Norovirus Hu/Norwalk/10255/2009/VNM
KC175380.1	2010	7511	Norovirus Hu/Norwalk/10285/2010/VNM
KC175381.1	2010	7511	Norovirus Hu/Norwalk/10296/2010/VNM
KC175382.1	2010	7511	Norovirus Hu/Norwalk/10313/2010/VNM
KC175383.1	2010	7511	Norovirus Hu/Norwalk/10325/2010/VNM
KC175384.1	2010	7511	Norovirus Hu/Norwalk/10328/2010/VNM
KC175385.1	2010	7511	Norovirus Hu/Norwalk/10368/2010/VNM
KC175386.1	2010	7511	Norovirus Hu/Norwalk/10378/2010/VNM
KC175387.1	2010	7511	Norovirus Hu/Norwalk/10386/2010/VNM
KC175388.1	2009	7511	Norovirus Hu/Norwalk/20008/2009/VNM
KC175389.1	2009	7511	Norovirus Hu/Norwalk/20010/2009/VNM
KC175390.1	2009	7511	Norovirus Hu/Norwalk/20014/2009/VNM
KC175391.1	2009	7511	Norovirus Hu/Norwalk/20016/2009/VNM
KC175392.1	2009	7511	Norovirus Hu/Norwalk/20019/2009/VNM
KC175393.1	2009	7511	Norovirus Hu/Norwalk/20033/2009/VNM
KC175394.1	2009	7511	Norovirus Hu/Norwalk/20035/2009/VNM
KC175395.1	2009	7511	Norovirus Hu/Norwalk/20044/2009/VNM
KC175396.1	2009	7511	Norovirus Hu/Norwalk/20047/2009/VNM

KC175397.1	2009	7511	Norovirus Hu/Norwalk/20066/2009/VNM
KC175398.1	2009	7511	Norovirus Hu/Norwalk/20067/2009/VNM
KC175399.1	2009	7511	Norovirus Hu/Norwalk/20069/2009/VNM
KC175400.1	2009	7511	Norovirus Hu/Norwalk/20092/2009/VNM
KC175401.1	2009	7511	Norovirus Hu/Norwalk/20093/2009/VNM
KC175402.1	2009	7511	Norovirus Hu/Norwalk/20094/2009/VNM
KC175403.1	2009	7511	Norovirus Hu/Norwalk/20118/2009/VNM
KC175404.1	2009	7511	Norovirus Hu/Norwalk/20122/2009/VNM
KC175405.1	2009	7511	Norovirus Hu/Norwalk/20123/2009/VNM
KC175406.1	2009	7511	Norovirus Hu/Norwalk/20128/2009/VNM
KC175407.1	2009	7511	Norovirus Hu/Norwalk/20135/2009/VNM
KC175408.1	2009	7511	Norovirus Hu/Norwalk/20139/2009/VNM
KC175409.1	2009	7511	Norovirus Hu/Norwalk/20140/2009/VNM
KC175410.1	2009	7511	Norovirus Hu/Norwalk/20142/2009/VNM
KC517361.1	2012	7511	Norovirus Hu/GII-4/Taoyuan/CGMH51/2012/TW
KC517362.1	2012	7509	Norovirus Hu/GII-4/Taoyuan/CGMH52/2012/TW
KC517364.1	2012	7509	Norovirus Hu/GII-4/New Taipei/CGMH54/2012/TW
KC517365.1	2012	7509	Norovirus Hu/GII-4/Taoyuan/CGMH55/2012/TW
KC517368.1	2012	7509	Norovirus Hu/GII-4/New Taipei/CGMH58/2012/TW
KC517369.1	2012	7509	Norovirus Hu/GII-4/Taoyuan/CGMH59/2012/TW
KC517372.1	2012	7509	Norovirus Hu/GII-4/New Taipei/CGMH62/2012/TW
KC517376.1	2012	7509	Norovirus Hu/GII-4/Taoyuan/CGMH66/2012/TW
KC517377.1	2012	7509	Norovirus Hu/GII-4/Taoyuan/CGMH67/2012/TW
KC517378.1	2012	7509	Norovirus Hu/GII-4/New Taipei/CGMH68/2012/TW
KC576909.1	2011	7509	Norovirus Hu/GII.4/GP2411/2011/USA
KC576912.1	2011	7509	Norovirus Hu/GII.4/GP13111/2011/USA
KC631814.1	2011	7509	Norovirus Hu/GII.4/MI001/2011/USA
KC631827.1	2012	7575	Norovirus Hu/GII.4/Hong Kong/CUHK6080/2012/CHN
KC810020.1	2010	7559	Norovirus Hu/GII.4/B/GBR/2010 clone PxB240610
KC810021.1	2010	7538	Norovirus Hu/GII.4/B/GBR/2010 clone PxB040710
KC810022.1	2010	7546	Norovirus Hu/GII.4/B/GBR/2010 clone PxB190710
KC810023.1	2010	7545	Norovirus Hu/GII.4/B/GBR/2010 clone PxB230710
KC810024.1	2010	7016	Norovirus Hu/GII.4/B/GBR/2010 clone PxB230810
KC810025.1	2010	7558	Norovirus Hu/GII.4/B/GBR/2010 clone PxB270710
KC810026.1	2010	7560	Norovirus Hu/GII.4/A/GBR/2010 clone PxA280610
KC810027.1	2010	7557	Norovirus Hu/GII.4/A/GBR/2010 clone PxA040710
KC810028.1	2010	7557	Norovirus Hu/GII.4/A/GBR/2010 clone PxA120710
KC810029.1	2010	7219	Norovirus Hu/GII.4/C/GBR/2010 clone PxC270810,
KC810030.1	2010	7557	Norovirus Hu/GII.4/D/GBR/2010 clone PxD240810
KC810031.1	2010	7560	Norovirus Hu/GII.4/E/GBR/2010 clone PxE230710
KC810032.1	2010	7560	Norovirus Hu/GII.4/E/GBR/2010 clone PxE230610
KC894942.1	2011	7558	Norovirus Hu/GII.4/Guangzhou/GZ2010-L88/CHN/2011
KC894943.1	2011	7556	Norovirus Hu/GII.4/Guangzhou/GZ2010-L91/CHN/2011
X86557.1*	1993	7556	Lordsdale virus

Appendix D: Robustness of the Norovirus Transmission Model to Genetic Clustering Threshold

D.1 2009-10 Season

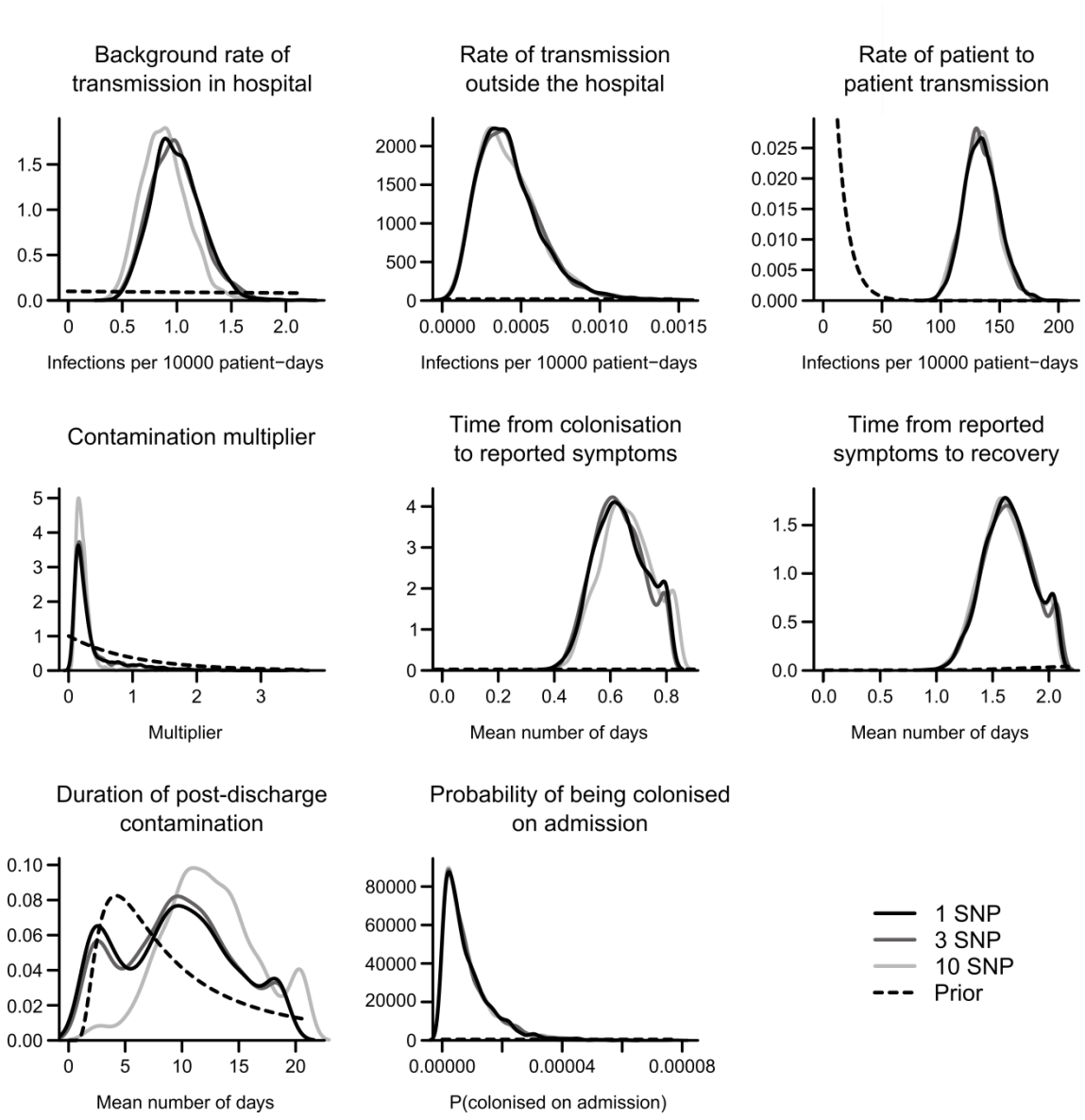


Figure D.1.1: Comparison of parameter distributions of the transmission model for 2009-10. Marginal posterior distributions for all parameters in the analysis; the prior distribution is shown with a dash black line, and the posteriors for the 1, 3, and 10 SNP thresholds shown in grey as marked.

D.2 2010-11 Season

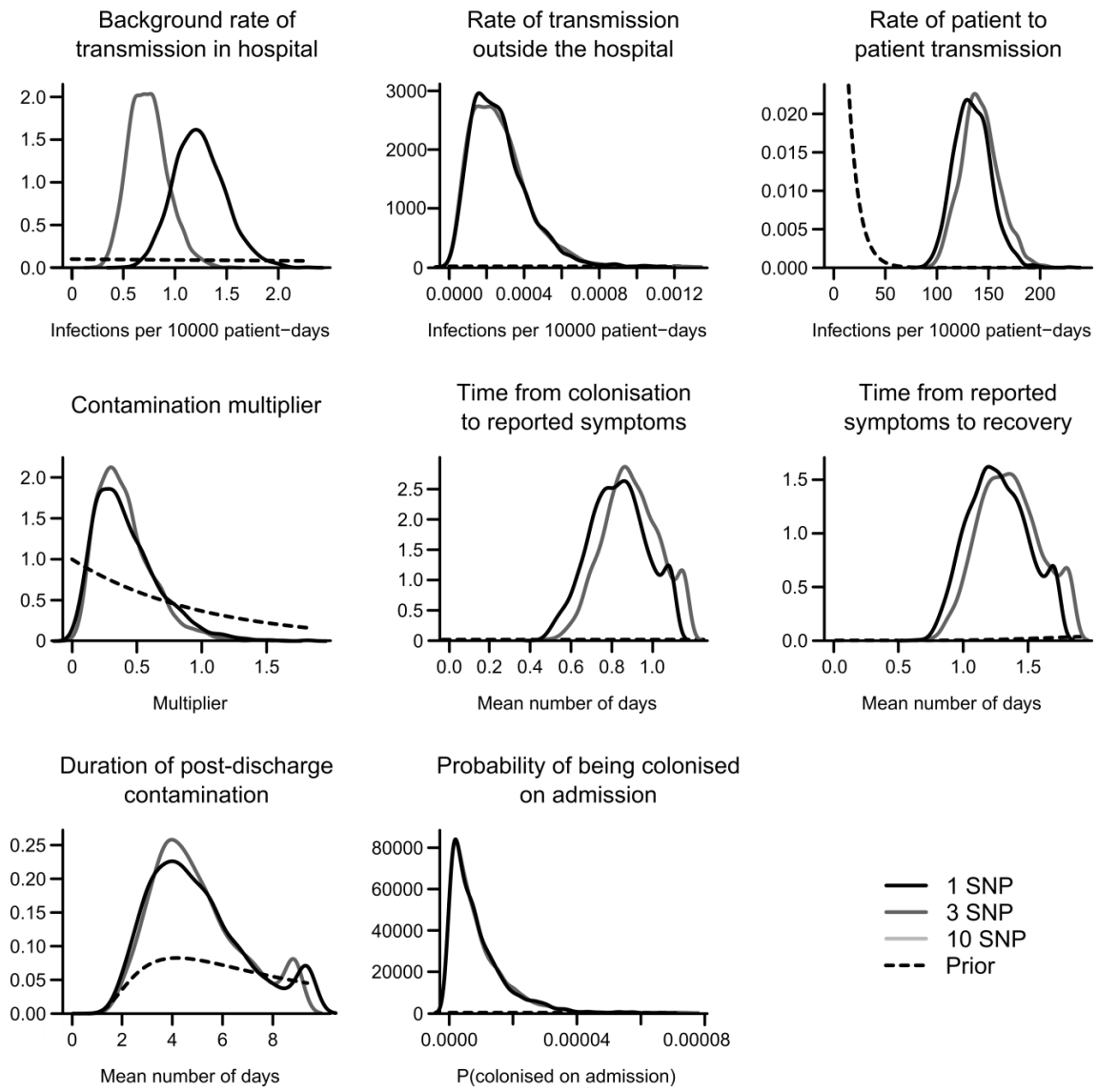


Figure D.2.1: Comparison of parameter distributions of the transmission model for 2010-11. Marginal posterior distributions for all parameters in the analysis; the prior distribution is shown with a dash black line, and the posteriors for the 1, 3 and 10 SNP thresholds shown in grey as marked. The SNP clustering is the same under the 3 and 10 SNP thresholds, and thus the lines are identical in this case.

D.3 2011-12 Season

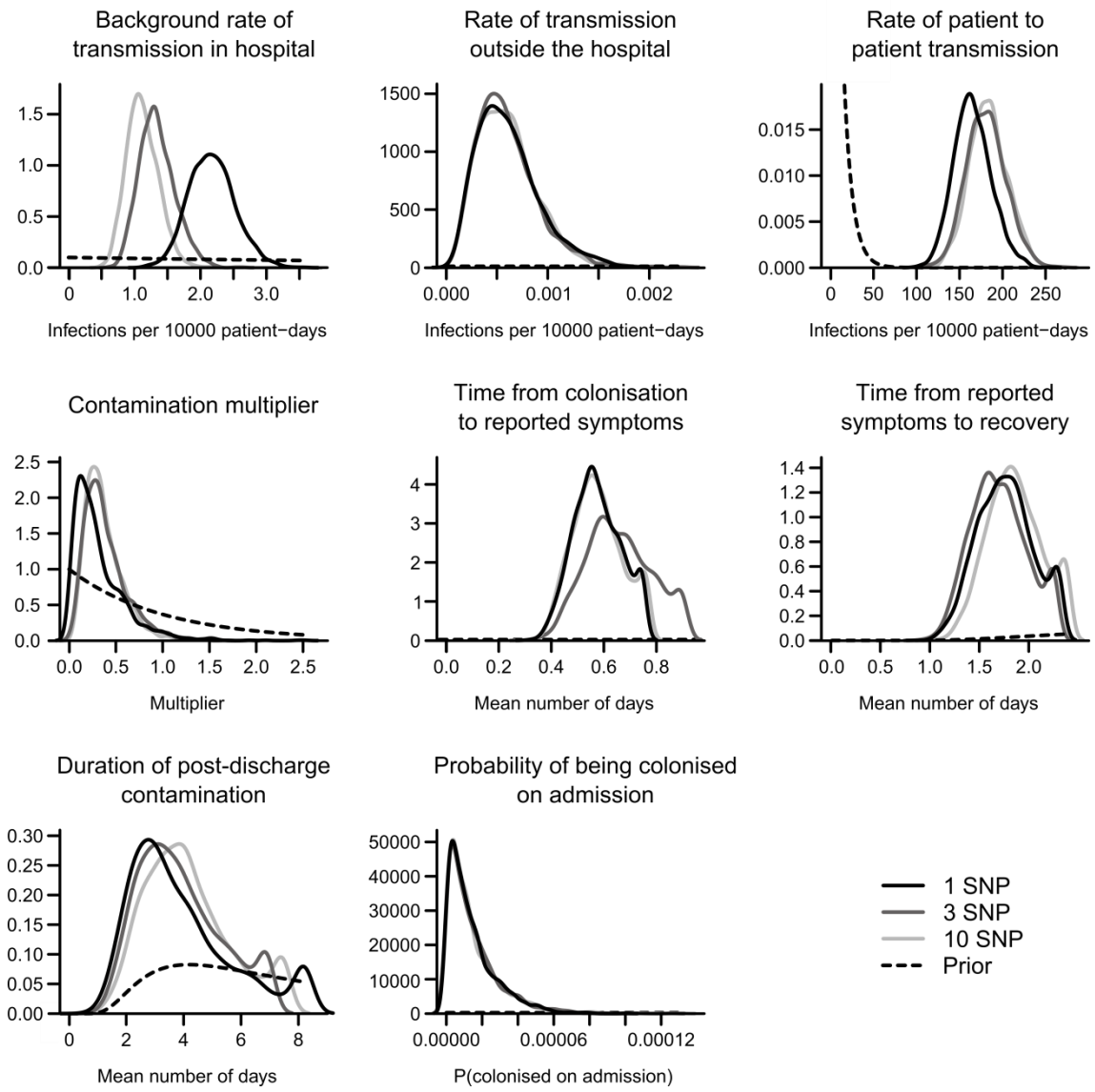


Figure D.3.1: Comparison of parameter distributions of the transmission model for 2011-12. Marginal posterior distributions for all parameters in the analysis; the prior distribution is shown with a dash black line, and the posteriors for the 1, 3 and 10 SNP thresholds shown in grey as marked.

D.4 2012-13 Season

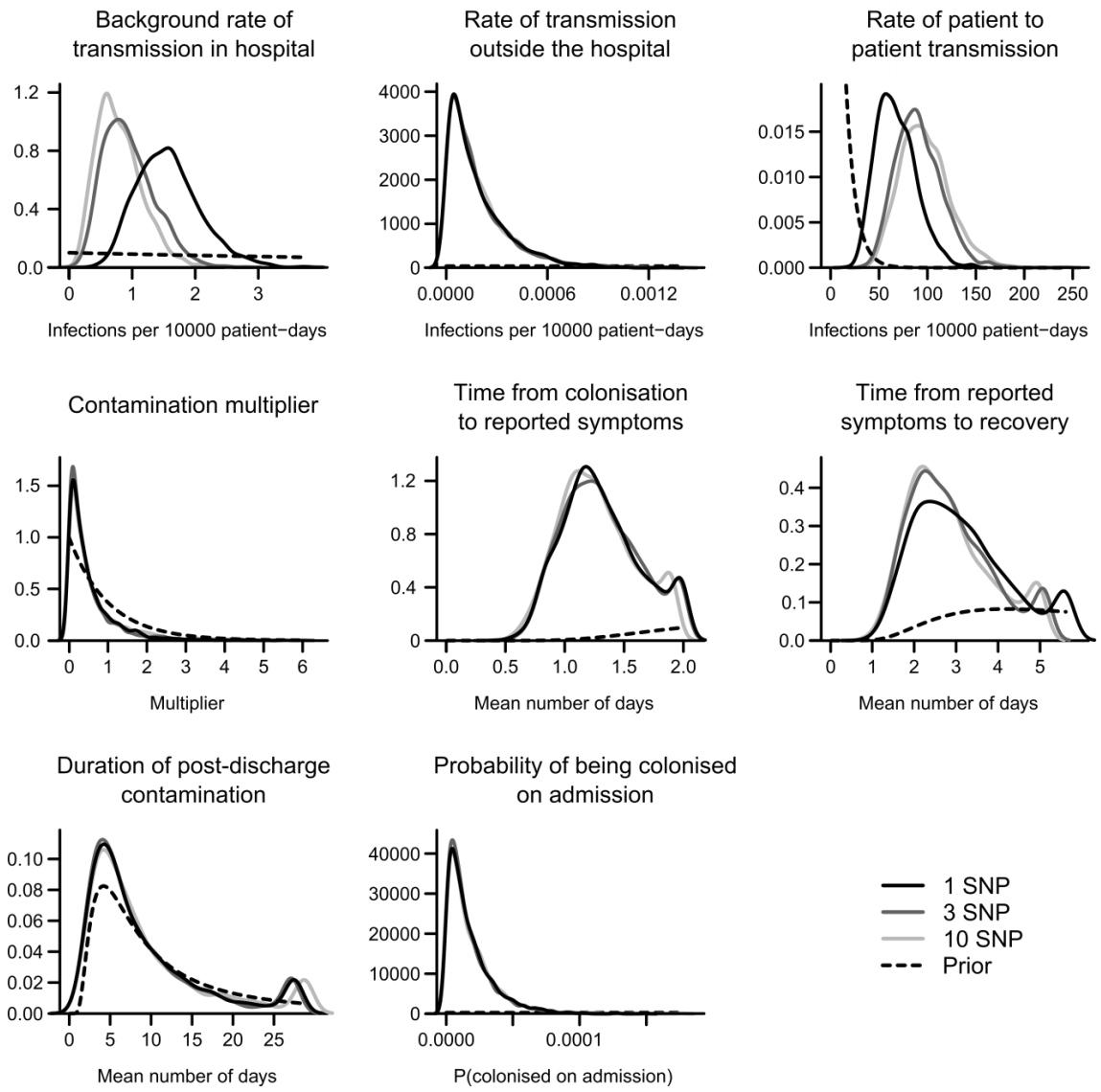


Figure D.4.1: Comparison of parameter distributions of the transmission model for 2012-13. Marginal posterior distributions for all parameters in the analysis; the prior distribution is shown with a dash black line, and the posteriors for the 1, 3 and 10 SNP thresholds shown in grey as marked.

Appendix E: *Campylobacter* Sequences

The sequences used in Chapter 5 are the same as those given in Table S1 from Sheppard et al. (2013b), with the following isolate numbers:

ST-21: 34, 36, 37, 40, 59, 60, 62, 63, 65, 71-78, 89, 94, 97, 110, 113, 116, 117, 182, 189, 195, 202, 203, 211

ST-45: 45, 48, 52, 55, 56-57, 70, 79, 81, 82, 84, 90-92, 100, 102-104, 111, 112, 114, 119, 124, 128, 131

ST-828: 2, 5, 15, 17-19, 21, 24, 98, 132, 134, 136-157, 159-163, 165, 167-171