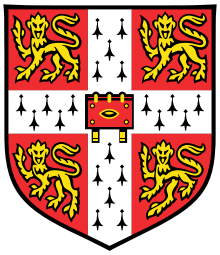
Inferring Determinants of Viral Transmission using Short-Read Sequence Data



Casper Kaalø Lumby

St Catharine's College
Department of Genetics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

January 2019

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the word limit of 60,000 words as specified by the Degree Committee for the Faculty of Biology.

Casper K. Lumby
Cambridge, January 2019

Summary

Title: Inferring Determinants of Viral Transmission using Short-Read Sequence Data

Author: Casper Kaalø Lumby

Summary: In order to spread, pathogens must not only be able to grow within an infected host, but also transmit to found new infections. In this thesis, I present a new population genetic framework generating insights into viral transmission events based upon genome sequence data collected before and after transmission. Previous attempts at bottleneck estimation have neglected the underlying genetic structure of viruses, considering instead less informative single-locus statistics.

Here I examine the problem of constructing reliable haplotypes from short-read sequence data, considering the performance of both exhaustive and minimal approaches in capturing linkage characteristics of the viral population. I present a simple method for bottleneck inference rooted in a multi-locus context supported by haplotype inference.

I next develop this model to incorporate selection for increased transmissibility, the effects of within-host growth, and noise arising from the sequencing process. Central to the method is a probabilistic model where unknown variables are marginalised over using compound distributions. A maximum likelihood scheme is employed in model selection where a machine-learning approach, referred to as adaptive BIC, was invented for the interpretation of likelihood statistics. I rigorously validate the performance of my model, identifying regimes wherein selection inference is feasible, and benchmark it against current state-of-the-art bottleneck inference algorithms, demonstrating a higher degree of realism and specificity within my approach.

I next extend the transmission model to account for advanced aspects such as selection for within-host viral adaptation, constructing a more realistic description of within-host growth processes. Accounting for within-host selection, I apply my transmission model to an experimental influenza transmission dataset in ferrets, providing novel quantitative insights.

I further explore limitations inherent to my model and consider regimes wherein the neutral version of my algorithm may be applied. I define and infer effective within-host selection for an influenza transmission study in pigs, employing my model to deduce a generally narrow transmission bottleneck in these animals.

Finally, I consider an influenza human challenge study and compute an effective single-segment within-host selection profile on the basis of an existing multi-segment characterisation. I discuss the relationship between human challenge studies and influenza infections occurring in a natural context.

Til min familie,

Cebrina, Christian, mor og far

Acknowledgements

My motivation for undertaking a PhD in bioinformatics stems from my time in Edinburgh where Davide, my Master's project supervisor, introduced me to the wonderful world that is biology. I am very grateful for the introduction and for your extreme patience when I failed to understand even the simplest of biological concepts.

During my time in Cambridge I have met many incredible people who have made it the most amazing time. I have had a number of fantastic co-workers and friends from within the Illingworth Group and the department in general: Mahan, Lei, Nuno, Ágnes, Saqib, Dominique, Bianca, Avinash, Tariq, Ruoyun and Sam. A special thank you goes out to Mahan, who, over the course of the year he was here, became a great friend of mine. You have one of the most positive views on life and exhibit a great affection for evolutionary biology, both of which I have found highly inspirational. I would also like to thank the many MGM students who have infiltrated Genetics over the years: Marcus, Lena, Nick, Marilou and Hilde. The department is slowly becoming more and more computational, which is a fantastic development. Of course, a massive thanks goes out to the Wellcome Trust and the various MGM programme directors for providing me with the opportunity for undertaking this PhD.

Over the years St Catharine's College has provided an excellent academic and social community for me to partake in outside of work. Throughout my almost five years here, I have established many invaluable friendships through college. I remember fondly my first year in Cambridge where the Russell Street Top Floor Crew made it the most amazing year and helped ease the transition into PhD life. Amongst others, I thank: Dave, Erin, Izzy, Will, Jonny, Ravi, Marcus, Michael, and Adam. A special thank you also goes out to St Catharine's College for bringing us all together.

Of course, a substantial proportion of my time in Cambridge has been spent in the company of boaties. Rowing for Catz has been a central part of my student life and keeping active has had a distinctly positive impact on my PhD work. Additionally, the boat club has been an important source of social interactions and personal development. I have made countless friends throughout my time in SCCBC and to mention them all would be impossible. Yet, here is a small list of people, without whom my time here would not have been the same: Will, Basile, Stuart, Beth, Geoff, Chris Q, JMG, Matt, Joe and many more...

Recently my life has been enriched with much banter due to some special people in Hot-Nerds 2.0: Jennie, Putu, Toby, Robyn and Stephen. I'm honoured of making the cut for the second edition of this group and thankful for the many laughs we have shared. Next year we'll definitely make it to Scotland. Maybe.

A special thank you also goes out to 4/5 of team WHEEL: James **W**agstaff, Toby **H**owison, Chris **E**ddy and Rory ⅀cMillan. When I first came to Cambridge, I vowed never to pull an all-nighter ever again. And I didn't. At least for the first three years. Then we decided to attempt a bike ride from Cambridge to Paris in under 24 hours, which is perhaps the most ridiculous thing I have ever done. Luckily, it prepared me well for when I had to break the vow a second time during the days leading up to my thesis deadline...

A quick thank you also goes out to my mates from Edinburgh: Ross, Joe, Dom, Sam and Ed. I'm forever grateful for your continued support and friendship.

The most heartfelt of thanks goes out to my family: Cebrina, Christian and mum. On occasion, these last two years have been the hardest I have ever had to endure. Thank you so much for being there for me. I feel that we, as a family, are now closer than ever. Finally, I thank my dad. Throughout my life you have been an inspiration to me. Everything I am today, I am because of you.

Lastly, but not least, I thank Chris. I couldn't have wished for a better supervisor. Despite the busiest of schedules, you always had time for me. You have supported me blindly throughout the last three years. Your constant optimism and excitement about biology is inspiring. I can't count the number of times where I have lost all hope in the project, thinking there is no way to solve a particular problem, then, after a quick chat with you, I am magically in possession of ten new ways of progressing. They may not always have been as mathematically rigorous as I would have wanted them, but they always moved the project forwards. Thank you so much.

Contents

List of Figures

List of Tables

List of Symbols

α	Dirichlet-multinomial noise coefficient for after transmission sampling, $\alpha = \frac{N^A+C}{1+C}$
β	Dirichlet-multinomial noise coefficient for before transmission sampling, $\beta = \frac{N^B+C}{1+C}$
δ	Coefficient relating to compound distribution, $\delta = \frac{N^T N^G - N^T - N^G + 1}{N^T N^G}$
γ	Coefficient relating to compound distribution, $\gamma = \left(\frac{N^T+N^G-1}{N^T N^G}\right)$
σ	Selection coefficient
$\boldsymbol{\mu}^B$	Mean of before population frequencies
Σ^B	Co-variance matrix for before population frequencies
C	Sampling noise
DS	Jacobian matrix with respect to the selection function S
g	Growth factor
$\boldsymbol{h}$	Set of haplotypes
L	Likelihood
$\boldsymbol{n}^A$	Number of copies of each haplotype in the after population
$\boldsymbol{n}^F$	Number of copies of each haplotype in the founder population
N^A	After transmission sampling depth
N^B	Before transmission sampling depth
N^G	Effective growth size

N^T	Bottleneck size
P	Probability
$\boldsymbol{q}^A$	Frequency of haplotypes in after transmission population
$\boldsymbol{q}^B$	Frequency of haplotypes in before transmission population
$\boldsymbol{q}^F$	Frequency of haplotypes in founder population
$S^G(.)$	Selection function for within-host adaptation
$S^T(.)$	Selection function for transmission
T	Matrix transforming full haplotypes to partial haplotypes
$\boldsymbol{w}^G$	Fitness for within-host adaptation of haplotypes $\boldsymbol{q}^F$
$\boldsymbol{w}^T$	Fitness for transmission of haplotypes $\boldsymbol{q}^B$
$\boldsymbol{x}^A$	Observation of haplotypes in after transmission population
$\boldsymbol{x}^B$	Observation of haplotypes in before transmission population

Chapter 1

Introduction

Viral transmission and within-host adaptation play key roles in the evolution and dispersion of viruses in a population. Due to a lack of proof-reading mechanisms, RNA viruses exhibit large mutation rates and represent ideal systems for studying evolution on short time scales. Determining how transmission and within-host growth each affect the adaptation of the virus is a crucial step towards appreciating the underlying mechanisms of viral evolution and may provide insights into development of antiviral procedures. Influenza viruses represent a highly studied and archetypal class of RNA viruses, having large mutation rates, short infection times and considerable virulence. In this work, I present a mathematical inference scheme for investigating viral transmission events, aiming on the one hand to separate evolutionary signatures due to transmission and within-host adaptation, and on the other hand to differentiate between stochastic and deterministic effects. I validate my inference framework on simulated data and apply it to datasets in ferrets, swine and humans. In this chapter I describe the biological and population genetic context surrounding my work.

1.1 Influenza Virus

Influenza A virus is an enveloped, single-stranded, negative-sense RNA virus. The viral genome is distributed across eight gene segments, each possessing a distinct purpose, as seen in Figure 1.1. Water fowl represent the natural reservoir for influenza, but the virus is known to infect a large range of hosts, including mice, pigs, humans, horses, and even whales (Haß et al. 2011; Kawaoka et al. 1998; Ma, Kahn and Richt 2008). During infection, influenza viruses cause inflammation of the host's respiratory system with varying degree of severity.

Seasonal influenza outbreaks are common and lead to mild infections which are typically cleared within a week. Regardless, seasonal flu can be fatal for the very young and the elderly, and it is estimated that up to 650,000 people die from influenza each year (Pagani et al. 2015; World Health Organization 2018b). Moreover, seasonal influenza has a considerable impact on economy through the loss of workforce and an increased need for medical attention (Tsai, Zhou and Kim 2014). Conversely, pandemic outbreaks are infrequent, but highly devastating. In the past one hundred years, four influenza pandemics (1918, 1957, 1968, and 2009) have caused great mortality within humans (Garten et al. 2009). In particular, this year marks the 100 year anniversary of the 1918-19 'Spanish flu' pandemic outbreak which coincided with the end of World War I and claimed the lives of an estimated 50-100 million people (N. P. A. S. Johnson et al. 2002). Recently, highly pathogenic avian viruses have been under scientific scrutiny due to evidence of sporadic transmission to humans; it is possible that the next big pandemic is of avian origin (Nature 2008; Peng et al. 2014; Watanabe et al. 2014; Wilker et al. 2013). In fact, the UK Cabinet Office identify influenza pandemics as one of two most severe national risks — the other being large scale attacks, including nuclear terrorism (UK Cabinet Office 2017).

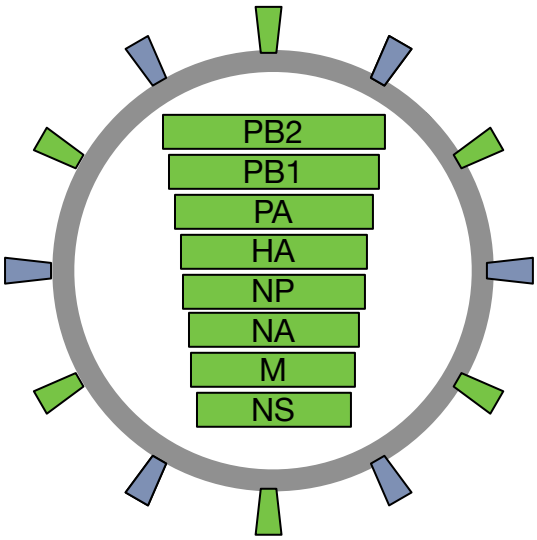


Figure 1.1. Diagram showing the eight RNA segments comprising the influenza genome.

Upon infection the host's immune system springs into action with the innate immune response initially leading the offensive. The innate immune system is unspecific in its nature, targeting the virus by producing a variety of cytokines, such as interferon, which obstructs viral replication and generates resistance in

neighbouring cells (Baccam et al. 2006; Pawelek et al. 2012). Whilst the innate immune system limits the infection, the virus is not eliminated until the onset of the adaptive immune response, which has been found to kick in about five days into infection (Miao et al. 2010; Pawelek et al. 2012). By generating an immunological memory, the adaptive immune system elicits virus-specific antibodies and T cells, which together clear the infection (Chen et al. 2018; Sandt, Kreijtz and Rimmelzwaan 2012). A high mutation rate allows influenza viruses to evolve and evade the host's immune response; natural selection acts in favour of variants conferring immune escape (Doud, Lee and Bloom 2018). In turn, this makes the development of influenza vaccines troublesome, with seasonal influenza vaccines having to be updated yearly to reflect antigenic drift (Carrat and Flahault 2007). Much effort has been dedicated to the development of a universal influenza vaccine, most recently with the Bill & Melinda Gates Foundation pledging up to $12 million in funding to promising research projects (The Bill & Melinda Gates Foundation 2018). Universal vaccines are generally aimed at targeting conserved regions within the virus, of which the so-called stem and globular head of the influenza haemagglutinin gene are prime candidates (Sautto, Kirchenbaum and Ross 2018). Haemagglutinin (HA) is one of two influenza surface glycoproteins, the other being neuraminidase (NA), and as a result makes for a good immune response target. Influenza nomenclature derives from the surface glycoproteins, with haemagglutinin (H) and neuraminidase (N) designating influenza A virus subtypes, e.g. H1N1 or H3N2 (Liu et al. 2009). Currently there are 18 different haemagglutinin subtypes and 11 different neuraminidase subtypes, together generating dozens of distinct viral strains (Centers for Disease Control and Prevention 2018). From a practical perspective, haemagglutinin is responsible for the attachment of the virion to sialic acid residues on the surface of the target cell, whilst NA cleaves the virus from the host cell during viral budding. Two types of sialic acid receptors exist, namely SAα2,3Gal and SAα2,6Gal receptors. The SAα2,6Gal receptor is primarily found in mammals whilst the SAα2,3Gal receptor pertains predominantly to birds, but may also be observed in the lower respiratory tract of humans (Kawaoka et al. 1998; Shinya et al. 2006). Generally, avian influenza viruses bind to SAα2,3Gal whilst human influenza viruses target SAα2,6Gal receptors, thus making the presence of specific sialic acids an important determinant of viral host range (Cauldwell et al. 2014).

Influenza viruses propagate via three distinct modes of transmission: contact transmission, including transmission via fomites, droplet transmission, which refers to the propagation of viral matter by large ($\geq 5\mu$m) respiratory particles,

and, finally, by aerosol transmission, which denotes the spread of virus via small ($< 5\mu$m) respiratory particles (Cowling et al. 2013; Gustin et al. 2011). Viral RNA has been found to survive in the air at distances of more than 100m away from the source (Scoizec et al. 2018). Influenza transmission has been studied extensively, both on local and epidemiological scales. On a local scale, the ferret model of transmission has long been the established approach for experimental influenza studies, owing to their small size, accurate emulation of human clinical conditions, and high identity to the human respiratory system (Belser, Katz and Tumpey 2011; Gustin et al. 2011). Transmission studies in ferrets and guinea pigs have identified varying transmission bottlenecks, signatures of natural selection, and the adaptation to and transmission of highly pathogenic avian viruses (Moncla et al. 2016; Varble et al. 2014; Watanabe et al. 2014; Wilker et al. 2013). Recently, transmission studies in humans have become increasingly prevalent, with data either deriving from natural infection (McCrone et al. 2018; Poon et al. 2016; Sobel Leonard et al. 2017b) or from direct challenge by the viral agent (Killingley et al. 2012; Sobel Leonard et al. 2016). Transmission to humans may also occur from animal hosts; zoonotic infections due to swine and poultry have been observed regularly, albeit without sustained transmission in humans (Bowman et al. 2017; World Health Organization 2018a). Whilst influenza viruses are believed to lack homologous recombination (Boni et al. 2008; Chare, Gould and Holmes 2003), which is a method whereby genetic material is exchanged between two strands of RNA, the segmented nature of their genome facilitates a reshuffling of genes, known as reassortment, which allows for the generation of novel viruses during replication in a multiply infected host cell. Reassortment is believed to have played a role in the emergence of viral pandemic strains transmitted to humans from swine (Ma, Kahn and Richt 2008; G. J. D. Smith et al. 2009).

On a global scale, influenza virus transmission patterns are shaped by the dissemination of epidemics from tropical regions, especially Southeast Asia, in which seasonal flu persists all year round (Hirve et al. 2016; Russell et al. 2008). The dispersal of influenza between continents is dominated by air travel (Lemey et al. 2014), whilst within-continent spread exhibits both radial patterns of spatial diffusion (Charu et al. 2017), as well as highly synchronised continent-wide outbreaks, with the specific propagation profile likely depending on the connectedness of the region (Geoghegan et al. 2018). Besides air travel, climate is an important driver of influenza epidemics in temperate regions, where seasonal reductions in absolute humidity result in increased rates of transmission and

heightened viral survival (Shaman and Kohn 2009). Globally, influenza evolution is governed by antigenic drift, i.e. mutations arising in genomic regions coding for antibody binding sites. Antigenic drift allows the virus to evade the host's immune system, in turn forcing the immune system to update its immune response; this continued arms race results in a rapidly evolving population and an influenza strain tree with a distinctive spindly structure (Łuksza and Lässig 2014). Strikingly, seven single amino acid substitutions in receptor binding regions have been found to collectively shape the majority of influenza antigenic evolution (Koel et al. 2013). It has also been shown that the combination of beneficial and deleterious mutations, i.e. not just beneficial mutations alone, have an important impact on evolution (Koelle and Rasmussen 2015); the mutational background upon which a beneficial variant lands is of high importance. Finally, as influenza transmission is governed by stochastic effects, within-host adaptation and viral transmission don't represent diverging processes (McCrone et al. 2018). Furthermore, it has recently been shown that within-host adaptation in immunocompromised individuals to a large degree mirrors global influenza evolution (Xue et al. 2017). This supports the idea that within-host adaptation is the main driver of global evolution in influenza. If true for influenza, this parallelism is likely not true for all RNA viruses. For the human immunodeficiency virus (HIV), the operations of transmission and within-host adaptations are misaligned; in the short term, within-host evolution results in a less transmissible viral population (Lythgoe et al. 2017). The host-specificity of immune pressure in this virus leads to a balance between host-adaptation and reversion, the loss of variants acquired in a previous host shifting the population towards an ancestral, transmissible state (Zanini et al. 2015).

1.2 Population Genetics

This thesis makes use of a number of concepts in population genetic theory. Population genetics, which was arguably founded as a discipline by Sewall Wright, Ronald Fisher, and J. B. S. Haldane, is a mathematical framework describing in a quantitative manner how evolutionary forces shape an evolving population.

1.2.1 Genetic Drift

Genetic drift refers to the stochastic process whereby the genetic composition of a population is changed over time due to the finite size of a biological population.

In so far as the next generation of a population is produced from a finite number of parents, the extent to which a genetic variant is present in individuals in a population will fluctuate over time (Charlesworth 2009).

Perhaps the most straightforward framework for the mathematical modelling of genetic drift is given by the Wright-Fisher model, as applied to a selectively neutral, asexual population. Under this model, each generation of individuals in the population is assumed to be distinct, instantaneously replacing the previous generation in a single moment in time. Each individual in the population has a unique parent, and itself has an equal probability of generating offspring, such that each subsequent generation may be considered as a multinomial sample drawn from that which came before it. Under this model we may consider the evolution of the frequency of a single variant in the population.

Given a population with N individuals we consider a biallelic locus with alleles a and A and suppose that the frequency of the allele A is given by p. Then in the next generation the probability of observing exactly k copies of the allele A is given by the binomial distribution:

$$\binom{N}{k} p^k (1-p)^{N-k} \tag{1.1}$$

Genetic drift becomes of increasing importance in small populations, where random changes are larger in magnitude. One scenario often considered is that of a population bottleneck, where a small sample of individuals found a new population. In this sudden decrease in population size, significant changes may be observed in the genotypic composition of the population.

A simple application of genetic drift has been in the analysis of population bottlenecks (Poon et al. 2016; Sobel Leonard et al. 2017a). We take the Wright-Fisher model as a model of viral transmission, in which a sample from a donor viral population gives rise to a new population. From the basic properties of a binomial distribution, we know that the expected frequency of the allele A is given by p, while the variance of A is given by $p(1-p)/N$. As such, where p^B and p^A are the observed frequencies of a variant before and after transmission, we have that

$$(p^A - p^B)^2 \approx \frac{p^B(1-p^B)}{N^T} \tag{1.2}$$

where N^T is the population bottleneck at transmission, and therefore

$$N^T \approx \frac{p^B(1-p^B)}{(p^A - p^B)^2} \tag{1.3}$$

In previous studies, where multiple alleles are present in a population, the assumption has been made that these alleles change in frequency independently of one another during transmission. In this thesis I examine this assumption in more detail exploring the consequences of the fact that alleles are joined together on chromosomes.

1.2.2 Linkage and Linkage Disequilibrium

Linkage is the name given to the fact that different alleles in a genome are physically linked together into a genome. Linkage was first proposed by Bateson, Punnet and Saunders, when they noticed that phenotypes in pea plants fell into non-Mendelian ratios, and suggested a physical coupling between genes (Griffiths et al. 2000).

Given variants at multiple positions in a genome we may consider a multi-locus variant, or haplotype. For example, given pairs of variant alleles (a,A), (b,B) and (c,C) at different positions in a chromosome, we may consider the haplotype aBc, describing an organism having these specific alleles in a chromosome or genomic segment.

Linkage disequilibrium describes the statistical association between variants at different positions in a genome. If within a population, having the allele a at a given genetic locus implies that there is greater or lesser probability of having the allele B at some other locus, then the loci are said to be in linkage disequilibrium. Linkage disequilibrium does not imply physical linkage; for example effects such as epistasis between alleles may lead to linkage disequilibrium.

In this thesis I make extensive use of the idea of a haplotype to describe how genetic variants are linked together in viral genomes. I show in multiple places that this approach, considering the physical structure of viral genomes, is of value to approaches which seek to understand viral evolution.

1.2.3 Selection and Epistasis

In an evolutionary process, selection may act upon the genomes in a population. Selection refers to the fact that organisms with different genomes may have, at the moment of conception, a differing probability of producing offspring in the next generation. Selection may act through a broad number of different phenotypes, including in a viral context the ability of a virus to withstand changes in temperature, the energetic landscape underlying protein folding, the ability to

bind host receptors, and the ability to evade the host immune system. In the bulk of this thesis, my concern will not explicitly be with viral phenotype, but rather in the fact of selection. We say that an allele is under selection if possession of that allele modifies the probability of individuals producing offspring; this probability may be represented in terms of the 'fitness' of the organism. Positive selection increases fitness, while negative selection decreases fitness.

Epistasis is a process whereby an allele modifies the fitness of an organism in a manner which depends upon the presence or absence of another allele, or multiple alleles, in the genome. Again, epistasis may be either positive or negative. Taken together, we define the fitness of an organism as a sum of single-locus and multi-locus, or epistatic effects. We may write

$$w = \sum_i s_i + \sum\nolimits_{i,j} \chi_{i,j} + \sum\nolimits_{i,j,k} \chi_{i,j,k} + \dots \tag{1.4}$$

where the fitness w is comprised of one-locus effects, s_i, dependent on the allele at locus i, of two-locus effects $\chi_{i,j}$, dependent upon interactions between alleles at the loci i and j, and so on.

Mathematically, the fitness of an organism may be converted into the probability that a specific individual will be the parent of an individual in the subsequent generation. Where w_i is the fitness of individual i, the probability that any given individual in the next generation is the offspring of i is given by

$$\frac{w_i}{\sum_j w_j} \tag{1.5}$$

where the sum is taken over the entire population.

The effect of genotype upon the fitness of an individual has been illustrated with the concept of a fitness landscape (Visser and Krug 2014). A fitness landscape provides a description of the fitness of an individual in terms of its genotype. In many cases, fitness landscapes are drawn, envisaging the space of all genotypes as a two-dimensional continuous space, with fitness represented by the height of a landscape covering that space. While somewhat tenuous in its representation of a discrete space, this conveys useful concepts such as fitness peaks, from which all changes to the genome lead to a decrease in fitness, and fitness valleys, from which a large number of directions of change increase the fitness of an organism.

1.2.4 Mutation

Mutation describes the process whereby the replication of genomes is prone to error. Whereas selection and drift each remove genetic variation from a population, favouring one variant over another, or stochastically changing frequencies until they hit one or zero, mutation introduces variation into a population, providing the genetic material upon which evolution can act.

Mutation rates may vary substantially across different organisms. Whereas studies of human populations have suggested a mutation rate close to 0.5×10^{-9} per year (Scally and Durbin 2012), a recent study of influenza virus identified a mutation rate of 1.8×10^{-4} per cellular generation (Pauly, Procario and Lauring 2017).

1.2.5 Recombination and Reassortment

Recombination is the process whereby genetic material from different 'parental' genomes is combined into new genetic material. In viral populations recombination is not necessarily an inherent part of reproduction, but may occur, for example through template switching, whereby the viral polymerase switches from one strand of RNA to another during replication. Within influenza populations, there is limited evidence for recombination occurring during the course of infection (Boni et al. 2008).

Reassortment is a process whereby parts of the genome not in genetic linkage with one another are shuffled during viral reproduction. In influenza this occurs during the process of intracellular reproduction. Different genetic segments of the virus are separated during reproduction, coming back together in the final production of a new virus. Because of this, where distinct viruses infect a single cell, genetic material from the two viruses can be shuffled, creating new combinations of virus. Reassortment is a key process in the formation of novel pandemic viral strains (G. J. D. Smith et al. 2009). Experiments conducted *in vitro* and in small mammalian systems have suggested an inherently high rate of reassortment (Marshall et al. 2013; Tao, Steel and Lowen 2014), though studies in human infection have suggested that spatial separation of genetically distinct viruses may lead to the effective rate of reassortment being much lower in these cases (Sobel Leonard et al. 2017a).

1.2.6 Evolutionary Change

The combination of mutation, selection, and genetic drift over time lead to evolutionary changes in populations. Changes of two distinct types may be observed. Over sufficiently long periods of time, substitutions may occur in a population, whereby a variant becomes fixed in the population. Such changes are often measured by changes in the viral consensus sequence; a measurement which can be conducted by sequencing a small number of viral genomes. Regarding HIV infection, phylogenetic methods have provided a good deal of insight into patterns of viral spread (Leitner et al. 1996). On very short timescales fixations in a population can be rare, albeit that changes on a smaller scale, such as those occurring in allele frequencies, can be exploited to study viral evolution. Population genetics, in providing a quantitative framework for the evaluation of such changes, is of great use in such situations (Illingworth 2015); population genetic theory is here used as the foundation for this thesis.

Chapter 2

Haplotype Reconstruction and Applications

2.1 Introduction

In this chapter I discuss in greater detail the principles of haplotype reconstruction. Given a viral population, described by short-read data, we define this as the inference of the underlying full-genome-length sequences which comprise the population, and the frequencies of these sequences within the population. For simplicity, haplotypes are often described in terms of the combinations of alleles which exist at polymorphic sites within the genome, all other alleles being preserved between haplotypes. The decomposition of populations into haplotypes, and changes occurring in haplotype frequencies, are key concepts which will be studied throughout this thesis. I here present studies utilising haplotype reconstruction in the context of three separate projects to which I have made partial contributions. These projects are further categorised into two themes: 1) haplotype inference in relation to within-host populations, and 2) haplotype inference with a view to transmission bottleneck estimation.

2.1.1 Author Contributions

The work presented in this chapter is currently unpublished. The work described here was carried out in collaboration with other researchers. The HIV fitness landscape project was conceptualised by Chris Illingworth and Matthew McKay with calculations of fitnesses being carried out by Saqib Sohail. The author contributed to the project by developing code for the reading of SAMFIRE

output (Illingworth 2015, 2016), inference of haplotypes and frequencies, and discussions on optimal approaches for inference. The work presented here focuses on the contributions made by the author.

The project on chronic influenza B infection was conceptualised by Chris Illingworth based on (currently unpublished) data obtained by Judy Breuer and collaborators. The population genetic analysis was a joint effort by Lei Zhao, Chris Illingworth and the author. The work presented focuses on the work conducted by the author, which forms part of a broader-ranging analysis of the data.

The transmission bottleneck inference project was conceptualised by Chris Illingworth and Daniel Weissman. Implementation of the method was carried out by Mahan Ghafari. The author contributed to model development, including mathematical derivations, generation of code for data simulation, and general discussions. The writeup is based upon an unpublished draft manuscript written in part by Chris Illingworth. Temporally, this work followed the development of techniques for evolutionary inference which are presented in Chapters 3 and 4. In this thesis the order of presentation is reversed as the approach described here is in many ways a simplification of what follows.

2.2 Haplotypes and Within-Host Populations

2.2.1 Introduction

A viral population may be defined by a set of haplotypes describing the specific alleles found at polymorphic loci within the genome. Haplotypes exist at specific frequencies in the population; these frequencies generally change over time as the population evolves. Upon sampling of the population, short-read sequence data may be produced in which the reads generally cover only a subset of the loci specified by the haplotypes. The problem of haplotype reconstruction considers the generation of such haplotypes, and the inference of their frequencies, on the basis of short-read sequence data. Accurate reconstruction of haplotypes is critical for reliable inference of within-host dynamics through the estimation of haplotype frequencies. In this section I present an exhaustive method for haplotype inference, developed by Illingworth (2015), and discuss its ability to infer the state of within-host populations through two projects. This approach to haplotype reconstruction separates the two stages of haplotype reconstruction, giving in the first place a list of haplotypes for which frequencies may be inferred

in a subsequent calculation. The first project considers the reconstruction of haplotypes and inference of frequencies in relation to understanding reversion dynamics in HIV. The second project considers the analysis of time-series data from an influenza B infection in a chronic patient. I here present mainly the findings related to inference of haplotype dynamics.

2.2.2 Methods

2.2.2.1 Exhaustive Method of Haplotype Reconstruction

Given a set of variant loci in the viral genome we may describe a viral population by a set of haplotypes, $\boldsymbol{h} = \{h_i\}$. A haplotype h_i is a sequence describing the specific alleles found at the polymorphic sites. When a haplotype describes all the variant sites in the genome, we refer to it as a *full haplotype*. At any one point in time, the set of haplotypes $\boldsymbol{h}$ is specified by an associated set of frequencies, $\boldsymbol{q} = \{q_i\}$, subject to the constraint that the frequencies must sum to unity, $\sum_i q_i = 1$. In the event that we were to sample and sequence the viral population, we would obtain observations $\boldsymbol{x} = \{x_i\}$ where x_i represents the number of times haplotype h_i was observed in the sample. This assumes that sequencing reads cover the entire genomic region specified by the haplotypes. Considering short-read sequencing approaches, this will generally not be the case, where, instead, reads span only a subset of the polymorphic sites. As such, short-read data cover only parts of the full haplotypes; we refer to subsets of full haplotypes as *partial haplotypes*. We here outline an exhaustive method for the reconstruction of the true haplotypes based on partial haplotype observations.

We utilise a haplotype reconstruction method developed by Illingworth (2015) and implemented in the SAMFIRE suite (Illingworth 2015, 2016); this code produces a list of haplotypes without inference of their frequencies. The reconstruction method produces full length haplotypes by merging multi-locus partial haplotypes where appropriate. In the event that all the partial haplotypes cover just a single locus, the reconstruction method necessarily generates 2^n haplotypes where n is the number of loci. Where partial haplotypes cover multiple loci, the number of reconstructed haplotypes is generally substantially smaller. The merging of partial haplotypes is based on three rules: 1) a redundancy rule, 2) an overlap rule, and 3) a combination rule. The redundancy rules leads to the removal of partial haplotypes fully contained within other partial haplotypes. The overlap rule merges overlapping haplotypes that report identical alleles in the overlap region. In haplotype reconstruction, rules 1 and 2 are repeated until

no further changes are observed. At this step, if partial haplotypes spanning $l < n$ loci still exists, a third rule, the combination rule, is employed. The combination rule generates all the potential full haplotypes for the $l < n$ partial haplotypes by combining them with the remaining haplotypes. Rigorous definitions of the reconstruction rules are given elsewhere (Illingworth 2015). This approach is exhaustive in the sense that it generates all possible full haplotypes from which the partial haplotypes could have been omitted. As a consequence, the set of reconstructed haplotypes will in general be greater than the true set of full haplotypes. Figure 2.1 demonstrates how reconstruction based on partial haplotype data from four full haplotypes results in seven potential haplotypes, of which the original four constitutes a subset.

2.2.2.2 Inference of Haplotype Frequencies

Given an observation of the viral population and a set of reconstructed haplotypes we next wish to infer the frequencies with which the haplotypes exist in the population. In the ideal world, viral samples may be described by a multinomial distribution in the sampling depth N and the frequencies $\boldsymbol{q}$. In reality, sampling is an imperfect process, and as such, an accurate representation must necessarily account for different sources of noise. To this end, we employ a Dirichlet-multinomial distribution with an overdispersion parameter C accounting for noise. Further description of the noise parameter C is outside the scope of this chapter; an in-depth discussion is provided in Chapter 3. The probability of obtaining an observation $\boldsymbol{x}$ from a set of frequencies $\boldsymbol{q}$ given sampling depth N is given by

$$P(\boldsymbol{x}|\boldsymbol{q}) = \frac{\Gamma(N+1)}{\prod_i (x_i+1)} \frac{\Gamma\left(\sum_i Cq_i\right)}{\Gamma\left(\sum_i x_i + Cq_i\right)} \prod_i \frac{\Gamma(x_i + Cq_i)}{\Gamma(Cq_i)} \tag{2.1}$$

where the index i denotes specific haplotypes. We note that Equation 2.1 assumes a full haplotype perspective; extension to a partial haplotype framework is straightforward, however, aiming here to provide a general overview of haplotype methods, we have deferred this to Section 2.3 and Chapter 3 in which an extensive description is provided.

Noting that Equation 2.1 describes not only the probability of the observations $\boldsymbol{x}$ given $\boldsymbol{q}$, but also the likelihood of the frequencies $\boldsymbol{q}$ given $\boldsymbol{x}$, we may obtain an estimate of the underlying frequencies by optimising this expression with respect to $\boldsymbol{q}$. In order to investigate the optimisation performance we inferred

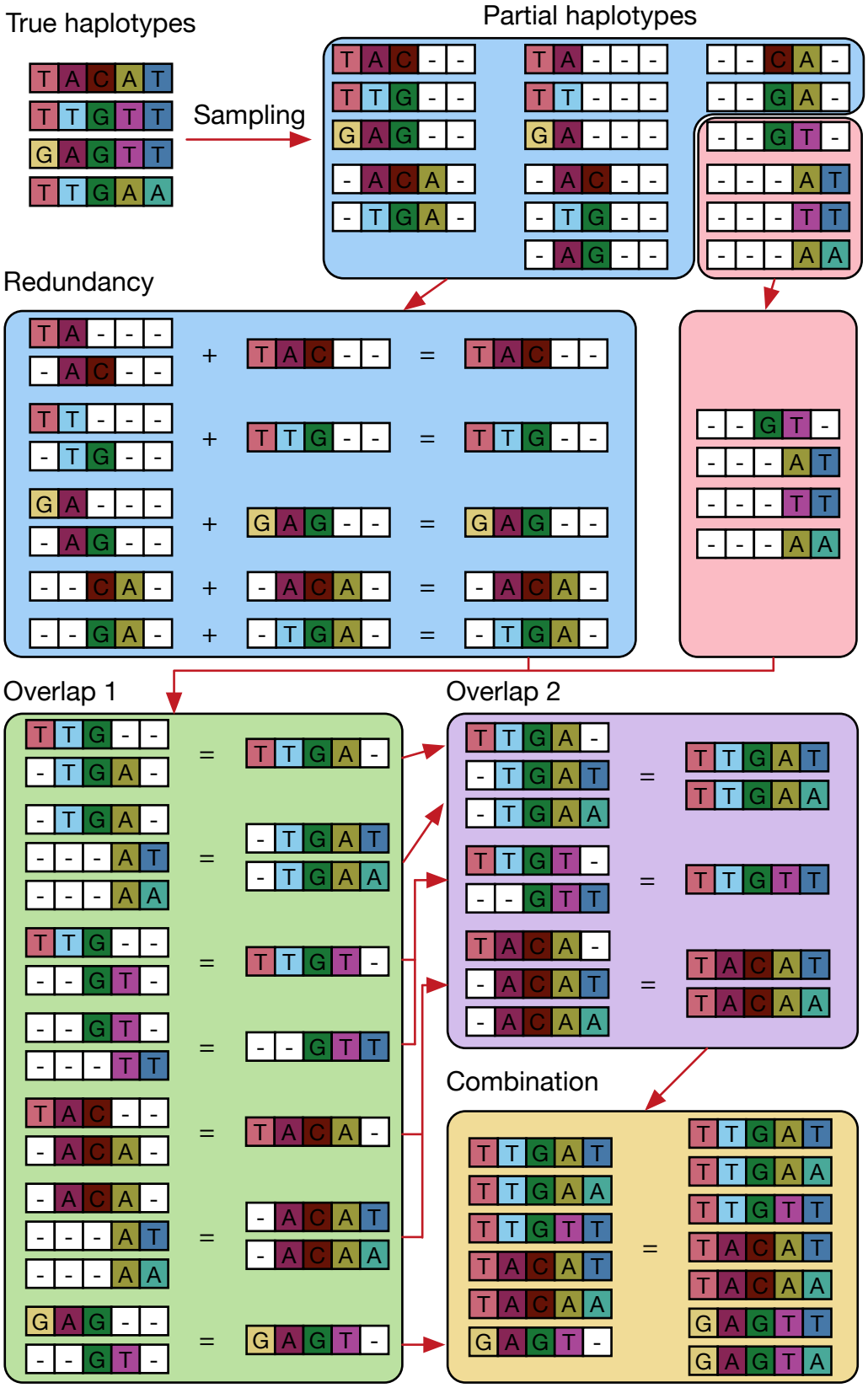


Figure 2.1. Illustration of the Illingworth (2015) haplotype reconstruction pipeline applied to sample data. (Continued on the following page.)

Figure 2.1. (Continued from previous page.) Four true haplotypes are sampled leading to 17 partial haplotypes of length $l \geq 2$ (single-locus partial haplotypes have been omitted for clarity). The redundancy and overlap rules are applied iteratively until no further changes come about. In the final step the combination rule generates the resulting seven reconstructed haplotypes.

haplotype frequencies from simulated data given a known set of haplotypes, i.e. without the additional complication of identifying which haplotypes exist in the population. Simulated data were designed to mimic sampling of short-read sequence data from five variant loci in the HA gene of the influenza A virus. Five different datasets were produced corresponding to $n_{\text{haps}} = \{2, 4, 8, 16, 32\}$ number of haplotypes, with each dataset spanning 200 simulation seeds. To avoid a situation wherein a subset of the haplotypes would substantially dominate the viral population, a heuristically derived minimum haplotype frequency of

$$q_{\text{min}} = \begin{cases} \frac{1}{n_{\text{haps}}+12} & \text{if } n_{\text{haps}} < 22 \\ 10^{-4} & \text{otherwise} \end{cases} \tag{2.2}$$

were employed. Considering the scenarios probed here, this resulted in minimum haplotype frequencies of $q_{\text{min}} = \{0.071, 0.063, 0.050, 0.036, 10^{-4}\}$ respectively. For the inference of haplotype frequencies we performed a maximum likelihood optimisation based around Equation 2.1, employing a simple hill climbing algorithm in the process. A full exposition of the simulation and inference framework is left for Chapter 3.

For comparison of optimisation performance we computed normalised Manhattan distances (ℓ_1-norms) defined as

$$d_1^{\text{norm}}(\boldsymbol{p}, \boldsymbol{q}) = \frac{\sum_i^k |p_i - q_i|}{k} \tag{2.3}$$

for frequencies $\boldsymbol{p} = \{p_1, p_2, \ldots, p_k\}$ and $\boldsymbol{q} = \{q_1, q_2, \ldots, q_k\}$.

Normalised Manhattan distances were computed between A) the inferred and the true frequencies and B) a random set of frequencies and the true frequencies.

2.2.3 Inference of Haplotype Frequencies

Results describing the performance of our method for frequency inference are shown in Figure 2.2. The left plot displays the distribution of normalised Manhattan distances in the case of eight haplotypes. We observe that the inferred frequencies are considerably closer to the true frequencies than a random set

of frequencies are. This shows the ability of the optimisation method to infer haplotype frequencies close to the actual values. The distances between the inferred and true frequencies have a low variance suggesting that the optimisation method performs well under most conditions.

The right plot displays the median normalised distance as a function of number of haplotypes. As the number of haplotypes are increased, both curves approach an asymptote of about 0.03. Median distances between the inferred and true frequencies are generally low, but increase marginally with number of haplotypes. This may be understood as the inability of the method to properly optimise the frequencies as the search space increases. Conversely, the distances between the random frequencies and the true frequencies diminish as the number of haplotypes is increased. This may be seen partly as the ℓ_1-norm's performance dropping as the dimensionality increases (Aggarwal, Hinneburg and Keim 2001) as well as a tightening of the range of values each entry in the random vector is likely to occupy. This doesn't necessarily mean that a random set of frequencies are on par with the optimised frequencies. For instance, as the number of haplotypes increase, the haplotypes become increasingly identical. As such, many of the partial haplotypes may be equally well represented by multiple full haplotypes; this results in an inference that may appear sub par with respect to distance from the true value, but which may still capture the partial haplotype characteristics well. On the other hand, whilst the random frequency vector might represent the frequencies of the majority of the dimensions well, it might entirely misspecify the frequency of a key haplotype representing several partial haplotypes.

2.2.4 Application 1: HIV Reversion Analysis

A recent approach to studying the evolution of HIV during the course of infection has been the application of matrix-based methods, which seek to describe a global fitness landscape for the virus. In short, an alignment is constructed of a very large number of HIV protein sequences, which together grant a representation of the global viral population. This alignment is converted into numerical values, assigning a '0' for the consensus amino acid, and a '1' for any other amino acid. A function is then constructed from the alignment, measuring the extent to which deviation from the global sequence consensus (represented by a zero at every position) imposes a fitness cost upon the virus. A simple one-locus version of this metric considers the likelihood of observing a '1' at any position in the

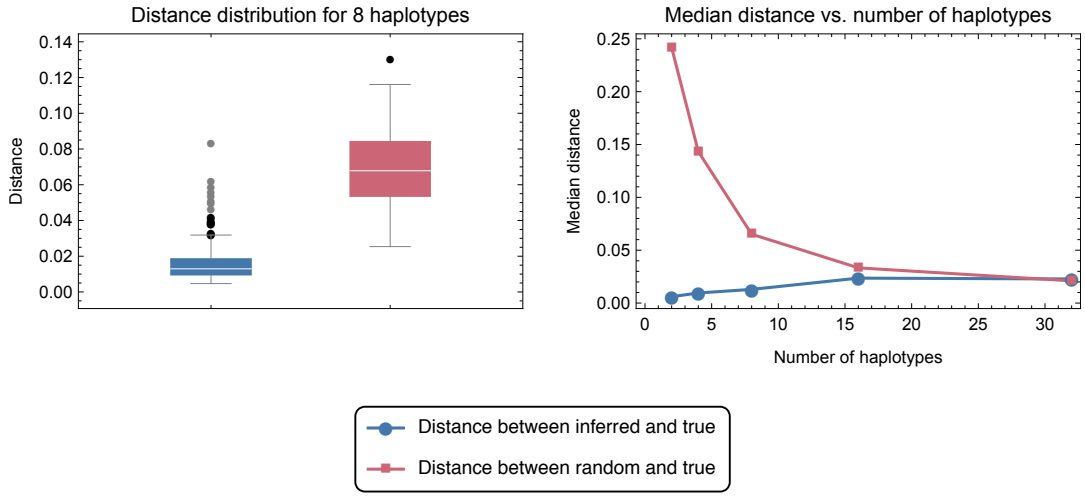


Figure 2.2. Comparison of normalised Manhattan distances from haplotype frequency inferences across 200 seeds. Two separate cases are shown: 1) distance between true and inferred haplotype frequencies and 2) distance between a random set of frequencies and the true set of frequencies. Left: Distribution of normalised distances for the specific case of eight haplotypes. Right: Median normalised distance for $n = \{2, 4, 8, 16, 32\}$ number of haplotypes.

alignment; if amino acids are perfectly uniformly distributed at a given position, it would be expected that close to $19/20 = 95\%$ of records in a given position would be represented by a '1'. By contrast, if an amino acid was perfectly preserved at a given locus, none of the records would be equal to a '1'. The measure thus gives an indication of sequence conservation, assigning a higher fitness cost to deviation from a more conserved position. In practice, both one- and two-locus fitness effects are used to generate a combined metric, sometimes described as an energy function, which represents the fitness of a given amino acid sequence (Barton et al. 2016; Ferguson et al. 2013; Louie et al. 2018). This function is of potential value in the design of vaccine therapies (Ferguson et al. 2013), and in predicting within-host HIV evolution.

Traditionally, the energy function is applied to consensus sequences, each sequence having a specific 'energy' or fitness. We here explore the potential for haplotype reconstruction to be used to evaluate viral fitness effects; reconstruction allows for the use of all of the data collected from short-read sampling, and therefore may give new insights into viral evolution.

Data from this project was collected from the paper of Zanini et al. (2015). This publicly available dataset spans nine untreated patients and covers 5–8 years of infection with each patient being sampled 6–12 times across the period.

The SAMFIRE software package (Illingworth 2016) was used to generate partial haplotype reads for this dataset, based upon loci at which a minor variant allele reached a frequency of at least 10% for at least one point during the course of infection.

Having defined potential haplotypes, two approaches were used to haplotype reconstruction. In the first, a simple maximum likelihood reconstruction was calculated, assigning frequencies to haplotypes. In a second approach, the Akaike information criterion (AIC) (Akaike 1974) was used to reduce the number of haplotypes with non-zero frequency. Here, we penalised haplotypes with frequencies larger than 10^{-8}, attempting to avoid overfitting of the data and thus obtaining a smaller set of haplotypes having non-zero frequencies. The AIC was defined as

$$\mathrm{AIC} = 2k - 2L \tag{2.4}$$

where k is the number of haplotypes with frequencies larger than $q = 10^{-8}$ and L is the likelihood of the system. Optimisation of both maximum likelihood and AIC expressions were achieved using a hill climbing algorithm, the specifics of which are given in Chapter 3. For comparison, haplotypes were generated using a random sampling approach, in which haplotype frequencies were produced using a uniform distribution before normalising the sum of the haplotypes to unity.

Reconstruction of haplotypes of fragment 1 of patient p2 at time point 3 resulted in a total of 36 full length haplotypes. We then inferred 100 sets of associated frequencies using the three inference methods. Array plots specifying the inferred haplotype frequencies in the 100 statistical replicates for maximum likelihood and AIC approaches are shown in Figures 2.3 and 2.4.

For both the maximum likelihood and AIC plots, we note a degree of variance across the statistical replicates, i.e. different starting points for the optimisation process result in slightly different inference outcomes. This matches our expectation from simulated data (Figure 2.2). However, patterns can be observed in the data; the maximum likelihood and AIC methods in general associate a similar set of frequencies to the reconstructed haplotypes.

We next evaluated the extent to which different haplotype reconstructions led to different values of the energy function. Where the energy of a sequence s is given by $E(s)$, we calculate the energy of a haplotype h, denoted $E(h)$, by combining the variants described at variable sites in the genome, as specified

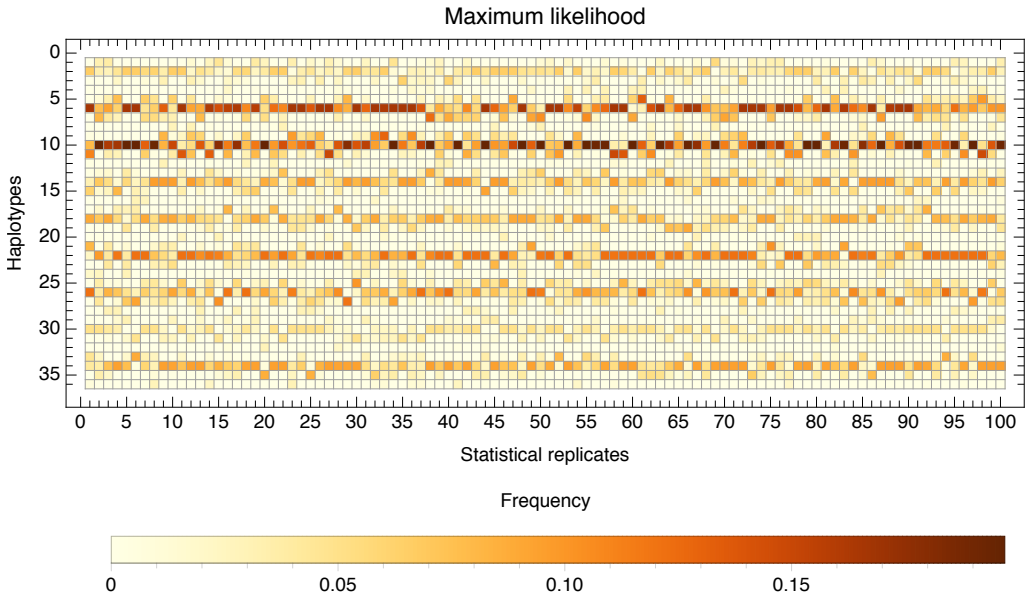


Figure 2.3. Haplotype frequencies for patient p2 from maximum likelihood inference. The y-axis denotes the 36 reconstructed haplotypes whilst the x-axis represents 100 statistical replicates. The colour of each grid tile represents the frequency of a haplotype.

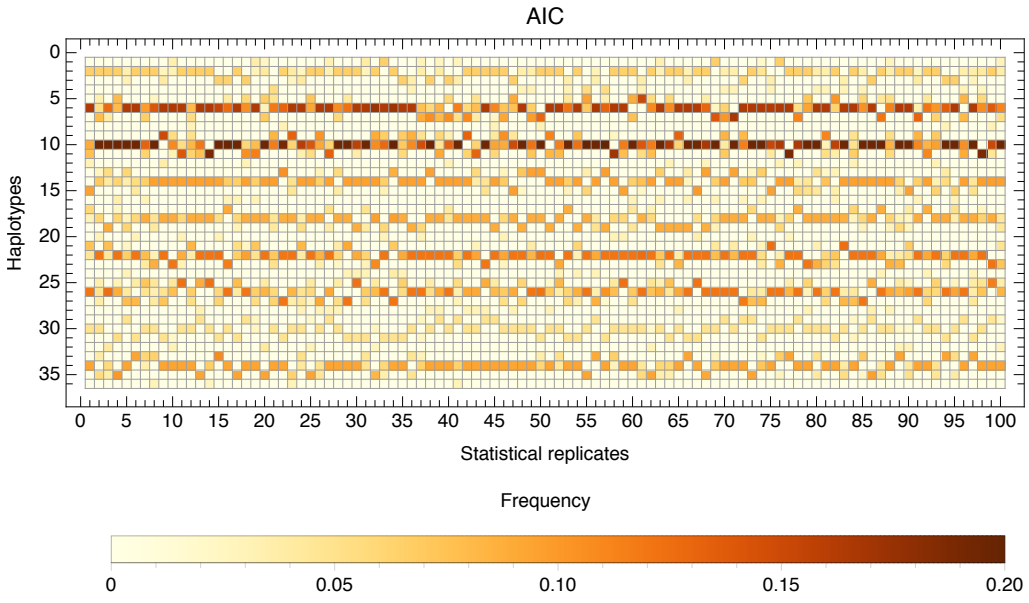


Figure 2.4. Haplotype frequencies for patient p2 from the AIC inference. The y-axis denotes the 36 reconstructed haplotypes whilst the x-axis represents 100 statistical replicates. The colour of each grid tile represents the frequency of a haplotype.

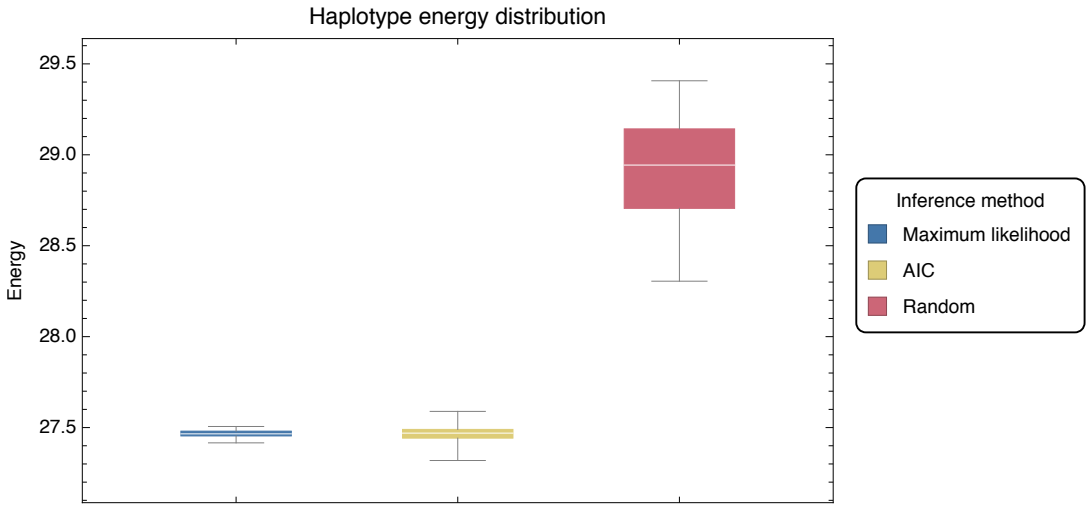


Figure 2.5. Distribution of haplotype energies for patient P2 at time point 3 based on frequencies inferred using the maximum likelihood method, the AIC method, or the random frequencies method.

by the haplotype, with consensus nucleotides at other loci in the genome, then translate this sequence into amino acids. For a haplotype reconstruction $\boldsymbol{h}$, in which the haplotype h_i has frequency q_i, we calculate the total energy as

$$E(\boldsymbol{h}) = \frac{\sum_i q_i E(h_i)}{\sum_i q_i} \tag{2.5}$$

where the summation variable i denotes the individual haplotypes in $\boldsymbol{h}$. We note that a low energy corresponds to a sequence which is closer to the overall consensus of the HIV sequence alignment, which indicates a higher fitness.

Outcomes of the weighted energy scores are shown in Figure 2.5. Values are shown for each of the 100 replicate sets of haplotype frequencies generated using the maximum likelihood, AIC, and random inference methods. The results show that while individual inferences of haplotype frequencies may differ, they are relatively strongly conserved with respect to the energy measure. This may be explained in a simple manner; while some degeneracy in the haplotype reconstruction exists, specific lower-dimensional properties of the haplotype frequencies, such as the frequencies of one- or two-locus variants in the dataset, are more highly preserved. This result suggests that haplotype reconstruction may be a viable method in calculating energy functions.

In a final analysis, we calculated changes in the within-host fitness of a HIV population using both the consensus and haplotype reconstruction methods. Results show firstly, that the inferred energy for the reconstructed population

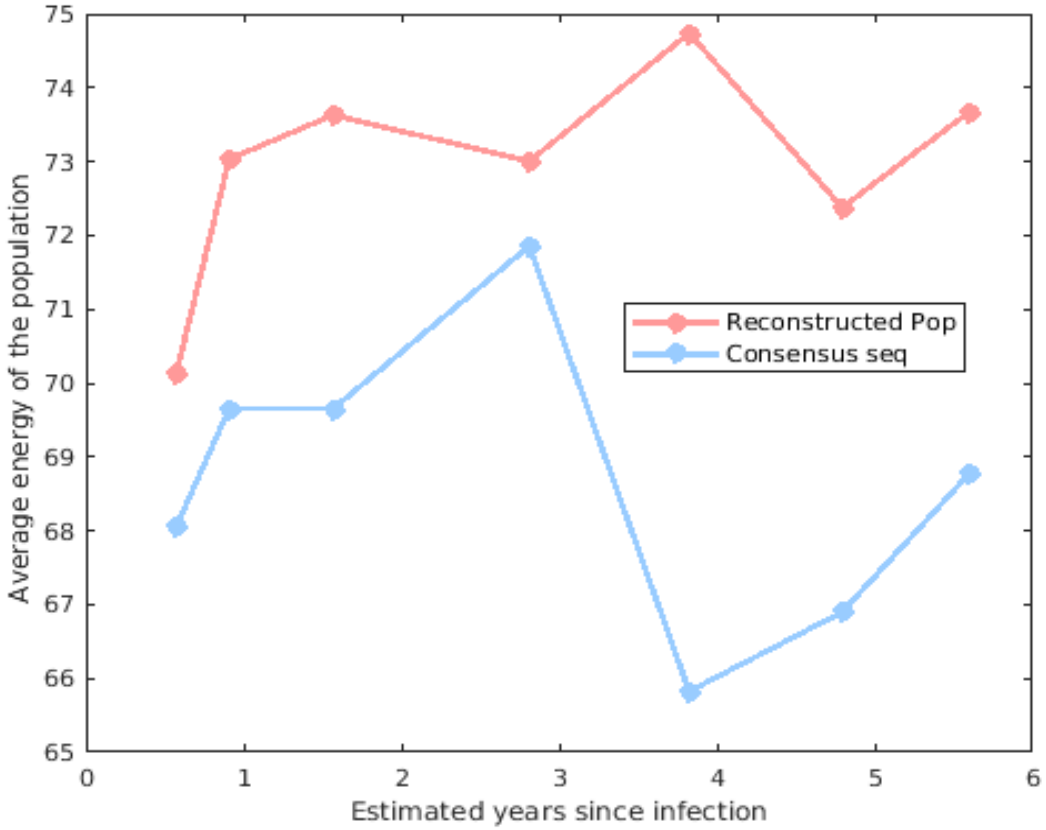


Figure 2.6. Energy values for patient P11 calculated using the consensus method and using data from the maximum likelihood haplotype reconstruction.

is higher than that calculated for the consensus (Figure 2.6). This is somewhat to be expected given the manner by which the statistic is calculated, indicating that the consensus of the within-host population is closer to the global consensus than the average sequence in the within-host population. Secondly, we note that changes in the statistic calculated by the reconstruction method do not always mirror those calculated from the consensus; a decrease in the consensus energy occurring in year three of infection is matched by an increase in the reconstructed population energy occurring at the same time point. We conclude that the consensus method of calculating energy values is potentially not the best approach to evaluating fitness within this alignment-based framework; use of short read data to generate a reconstructed population more faithfully represents the data, produces consistent results, and may differ substantially from the more traditional statistic.

2.2.5 Application 2: Haplotype Dynamics in Chronic Influenza B Infection

We applied the maximum likelihood method of haplotype reconstruction to analyse haplotype dynamics within a case of chronic infection with influenza B. Samples were collected at 41 time points from a child with a severe primary immunodeficiency. After initially failing to respond to treatment with neuraminidase inhibitors, the patient was treated with a combined zanamavir/favipiravir therapy. This led to the apparent clearance of infection for a period of one month, following which a sample showing a low positive result for influenza was collected. No action was taken until approximately one month after this, when the patient showed symptoms of respiratory infection. Samples collected at this time were positive for influenza infection. A second round of treatment with a combined zanamavir/favipiravir therapy again cured the infection, following which no further resurgence of infection was observed.

Data describing the course of infection are shown in Figure 2.7. The periods of initial infection and resurgence of infection can be clearly observed.

Viral sequencing was conducted, collecting data from each positive sample. Read depths from each sample are shown in Figure 2.8. Data from each sample were aligned to reference sequence data using BWA (H. Li et al. 2009) and further processed using the SAMFIRE software package (Illingworth 2016).

For each sample, a consensus sequence was calculated for each segment, joining the sequences for each segment into concatenated viral sequences. A Bayesian phylogeny was then constructed from these data using the BEAST 2 software package (Bouckaert et al. 2014) (Figure 2.9). This phylogeny highlights an internal structure within the population, which was divided into two clades, which we term A and B. While the majority of the sequences formed clade A, the two samples collected during the resurgent infection (samples 40 and 41) were distinct from these, clustering with a sample collected before favipiravir treatment (sample 18, indicated in Figure 2.7). The evolutionary consistency between the samples collected after the resurgence of infection and a prior viral population ruled out the possibility that the resurgence was caused by a re-infection with a new influenza virus.

In order to investigate the evolution of the population in greater depth, sequences from the neuraminidase (NA) segment were chosen for analysis. Drug resistance to two of the drugs dispensed to the patient, namely oseltamivir and zanamivir, is known to occur in this segment (Hurt et al. 2009; Zürcher et al.

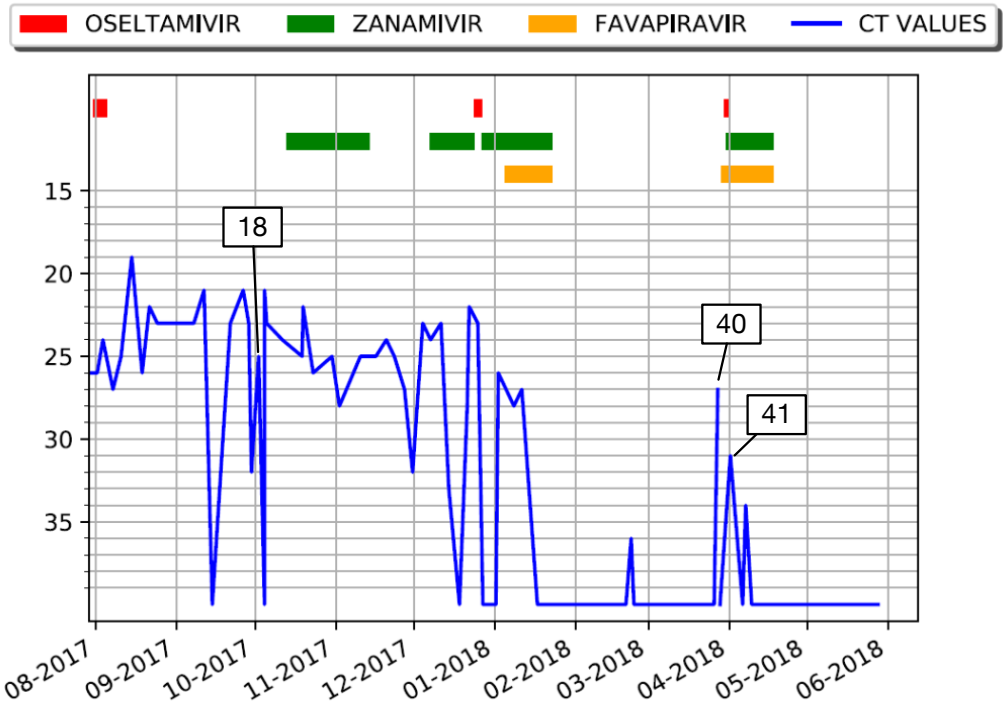


Figure 2.7. The blue line here shows CT scores, indicating the extent of infection within the patient. A low CT score indicates a higher viral load in a sample. A CT score of 40 indicates a negative test result. Bars show periods of administration of each drug. Selected samples are assigned numbers. The number 41 indicates the last positive sample to be sequenced.

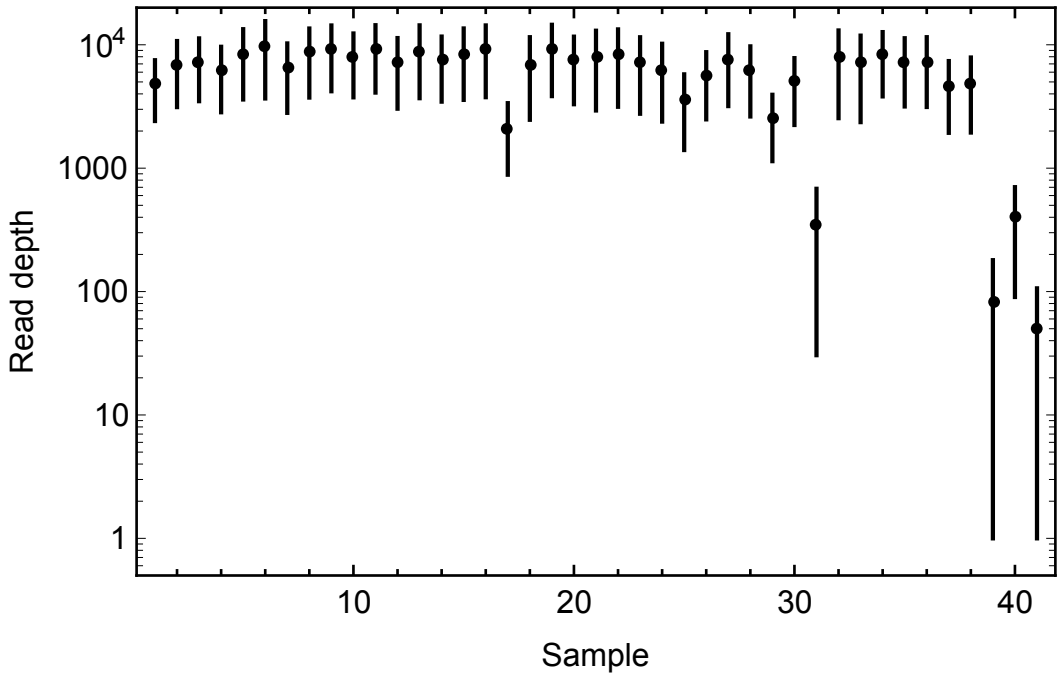


Figure 2.8. Mean and standard deviations of the read depths of samples collected from the influenza B patient. Where a standard deviation encompasses an interval crossing zero, the corresponding line is truncated.

2006); as such, this segment was chosen for analysis. Haplotypes were reconstructed across all time points using the SAMFIRE approach. Variant loci to be used in the haplotype reconstruction were identified as those which were fixed in the consensus sequences of samples 18, 40, and 41; that is, identifying loci at which substitutions were observed in the key branches of the tree. Next, multi-locus variants were called at these loci in the genome, constructing partial haplotype reads. Following this, a complete set of haplotypes were constructed using the method illustrated in Figure 2.1. Finally, the frequency of each haplotype were inferred at each time point using a hill climbing optimisation algorithm aiming to maximise the likelihood defined in Equation 2.1. Haplotypes inferred to reach a frequency of 10% or greater in at least one sample were identified to construct a visualisation of the haplotype reconstruction which had the maximum identified likelihood (Figure 2.10). As the reconstruction shows, haplotypes 13–15 and 17 were only observed in sample 18, while haplotypes 16 and 18–21 were only observed in samples 40 and 41. These haplotypes were never inferred to reach a substantial frequency in any of the other samples, indicating that this clade represents a distinct subset of samples within the viral infection. Examination of the haplotype data showed a single putative resist-

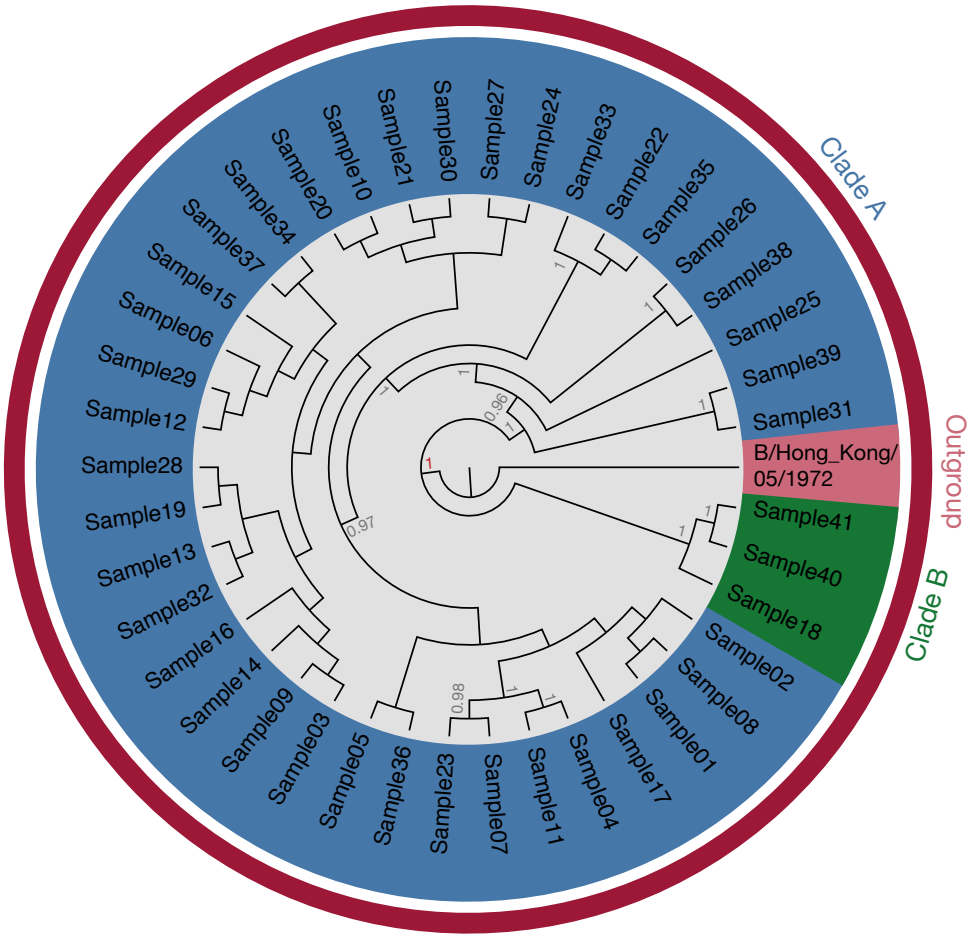


Figure 2.9. Bayesian phylogeny inferred from the influenza B patient data using B\Hong_Kong\05\1972 (GISAID sequence ID: EPI_ISL_1775) as a representative influenza B strain. Inference was conducted using BEAST 2 (Bouckaert et al. 2014) applying the HKY substitution model (Hasegawa, Kishino and Yano 1985). MCMC was run for a total of 10 million iterations with the first 10% of each run reserved for burn-in rounds. The target tree was generated using TreeAnnotator and graphics were produced with the aid of the FigTree software package (Rambaut n.d.). Within the tree we identify two distinct clades, coloured in blue and green respectively. Posterior levels of support greater than 0.95 are shown.

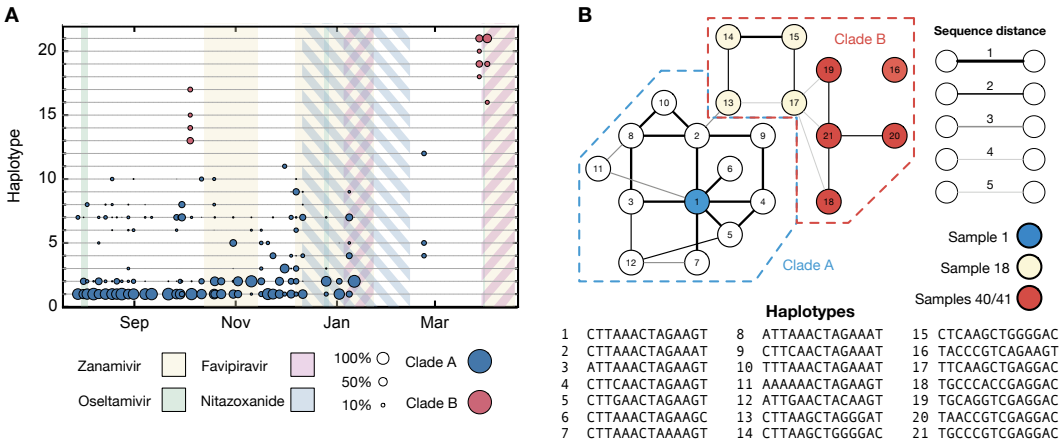


Figure 2.10. **A** Maximum likelihood haplotype reconstruction across the 41 sample points. In this plot, administered drugs are shown in green (oseltamivir), yellow (zanamivir), red (favipiravir), and blue (nitazoxanide). A total of 21 haplotypes were inferred to exist at a frequency of 10% or greater in at least one time point from the course of the infection; other haplotypes are neglected in this plot. Circles show the sizes of the respective haplotypes, having an area proportional to the frequency of the haplotype at each sample point; colours of circles indicate the clades to which they belong. **B** Overview of Hamming distance between haplotypes. Only the shortest distance between two haplotypes are shown.

ance mutation arising in the samples 40 and 41, in the residue E117. Where the variant E117D is known to induce drug resistance to zanamivir (Oh et al. 2018; B. J. Smith et al. 2001), we observed the substitution E117A. Whereas the E to D substitution results in the movement of a carboxylic acid group away from a position in which it binds the drug zanamivir, the E to A substitution removes this group entirely, suggesting that it has a similar effect in reducing the affinity of the drug for the NA receptor (Figure 2.11).

This analysis, while basic in nature, illustrates the potential of haplotype reconstruction methods to generate insights into sequence data. Our analysis here appears to show distinct viral populations within the host, potentially with genetically distinct viral clades at different locations within the host airway. Intra-host diversity of this nature has been identified in bacterial infections such as those suffered by cystic fibrosis patients (Lieberman et al. 2014), while genetic diversity in spatially separated influenza populations within a host has been found in a case post-mortem (Hamada et al. 2012). Such studies support the possibility of spatial diversity occurring in the patient from whom data were analysed here; a population existing in one part of the airway appears to have survived combined zanamivir and favipiravir therapy to go on to create a resurgent infection.

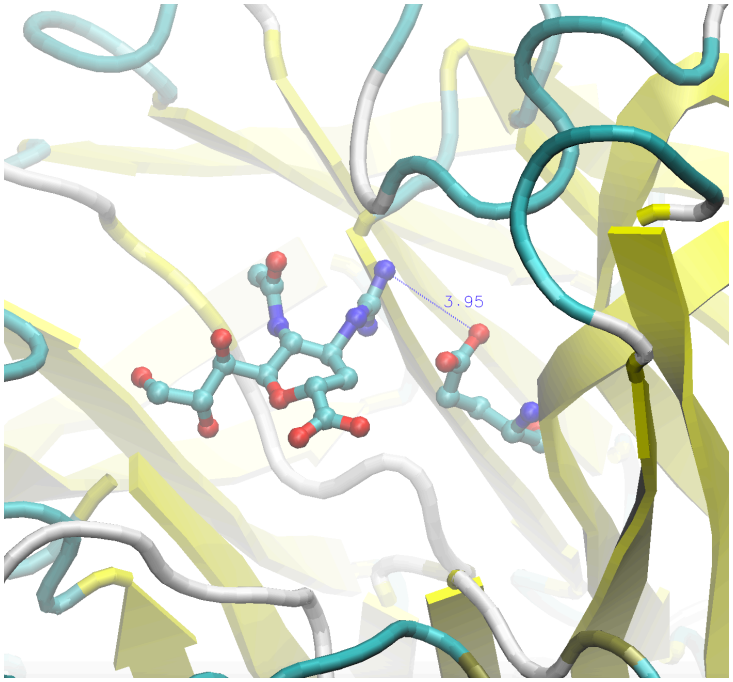


Figure 2.11. Binding interaction between the E117 amino acid in neuraminidase and zanamivir. The amino acid and the drug molecule are shown in ball and stick format; the potential hydrogen bond between the amino acid and the drug is highlighted. Mutation to aspartic acid would increase the distance between the two functional groups; the mutation we observe to alanine removes this interaction entirely. This image was created with the VMD software package (Humphrey, Dalke and Schulten 1996).

2.3 Haplotypes and Viral Transmission

2.3.1 Introduction

The approach described above for haplotype reconstruction describes a two-step, maximal approach to the inference of haplotypes. Firstly, a list of plausible haplotypes is drawn up. Secondly, frequencies for these haplotypes are inferred, either in a free manner, or constrained by some evolutionary model. Such a maximal approach has the advantage of describing a large number of potential haplotypes which may exist in a population, given the observed sequence data from one or multiple time points. In subsequent chapters, this property will be used for the inference of selection in viral populations; for a model to account for selection shifting the population between haplotypes, any potential haplotype into which the population can be shifted must be included in the model.

I here note that in cases where there is sparse partial haplotype data available to constrain the haplotype space and the number of variants is large, the set of haplotypes generated by this approach can become massive. In some cases the size of this haplotype space may preclude further calculations from being performed. We therefore investigated an alternative approach to haplotype reconstruction, which uses model selection to seek a minimally complex explanation for the dataset. We apply this approach to infer the bottleneck size underlying the transmission of influenza viruses using data from a household transmission study (McCrone et al. 2018).

2.3.2 Haplotype Inference Method

We constructed a maximum likelihood approach for haplotype reconstruction based upon existing technologies for processing short read data (Illingworth 2015, 2016; Illingworth et al. 2017). We here assume that we have short-read data describing a viral population both before and after a transmission event. In a preliminary step, an allele frequency cutoff was used to identify a list of polymorphisms which are present in the 'before transmission' data.

Given short-read sequence data, we next processed the reads into partial haplotypes, identifying sets of partial haplotypes as sets of reads which describe the alleles present at a consistent set of one or more polymorphic loci (Illingworth 2016). Each set of partial haplotypes represents an independent set of observations from the underlying viral population; the likelihood of this observation

given the underlying population may be described as a Dirichlet-multinomial distribution (Illingworth 2015). Partial haplotype data are collected for the samples obtained both before and after transmission; we denote these datasets as $\boldsymbol{x}_l^{B,P}$ and $\boldsymbol{x}_l^{A,P}$ respectively, where l denotes the partial haplotype set.

We now suppose that the viral population is comprised of a set of k haplotypes, $\boldsymbol{h} = \{h_1, h_2, \ldots, h_k\}$, with the frequencies $\boldsymbol{q}^B = \{q_i^B\}$ before transmission and $\boldsymbol{q}^A = \{q_i^A\}$ after transmission. These frequencies can be converted into partial haplotype frequencies by projection of the full haplotype space onto each lower-dimensional partial haplotype space by means of matrices T_l. For example, given the full haplotypes before transmission $\{\text{GA}, \text{TA}, \text{GC}, \text{TC}\}$ and a set of partial haplotypes $\{\text{G-}, \text{T-}\}$, we may write

$$\boldsymbol{q}_l^{B,P} = T_l \boldsymbol{q}^B \tag{2.6}$$

or more explicitly,

$$\begin{pmatrix} q_{l,1}^{B,P} \\ q_{l,2}^{B,P} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} q_1^B \\ q_2^B \\ q_3^B \\ q_4^B \end{pmatrix} \tag{2.7}$$

In this way we can construct a likelihood for the set of haplotypes and frequencies, for example

$$\log L(\boldsymbol{h}) = \sum_{t \in \{B,A\}} \sum_l \log L_D(\boldsymbol{x}_l^{t,P} | T_l \boldsymbol{q}^t, C) \tag{2.8}$$

where L_D denotes the Dirichlet-multinomial likelihood

$$L_D(\boldsymbol{x}|\boldsymbol{q}, C) = \frac{\Gamma(N+1)}{\prod_i (x_i+1)} \frac{\Gamma\left(\sum C q_i\right)}{\Gamma\left(\sum x_i + C q_i\right)} \prod_i \frac{\Gamma(x_i + C q_i)}{\Gamma(C q_i)} \tag{2.9}$$

in which $N = \sum_i x_i$ and C is a noise parameter which characterises the extent of information provided by the data $\boldsymbol{x}$ (Illingworth et al. 2017).

To construct the set $\boldsymbol{h}$, a set of k haplotypes are created one by one, randomly selecting partial haplotypes and joining them until k haplotypes, each spanning all polymorphic loci, are constructed. Following this, the inferred frequencies $\{q_i^*\}$ of the haplotypes are optimised to fit the data, under the constraint that $\sum_{i=1}^k q_i^* \geq 0.99$. This constraint allows for the inclusion of an additional, 'cloud'

haplotype, representing all of the haplotypes not included in $\boldsymbol{h}$; any partial haplotypes which do not match any of the specified full length haplotypes are assumed to be emitted from this final haplotype. A given set of haplotypes is required to explain the data collected both before and after the transmission event; while the haplotype frequencies are allowed to differ between these datasets, the intrinsic set of haplotypes was preserved across transmission. Optimisation of frequencies were performed using a hill climbing algorithm with a heuristically derived convergence criteria defined either as the absence of an improvement in likelihood across 50 consecutive frequency updates or as the completion of 8000 update iterations overall.

Optimisation of $\boldsymbol{h}$ was conducted by making perturbations to haplotypes in this set, calculating a maximum likelihood set of frequencies at each instance. Iterating this process identified the optimal set of k haplotypes given the data. Repeating this for increasing values of k gives a series of fits to the data; we use the Bayesian Information Criterion (Schwarz 1978) to identify the most parsimonious explanation for the data:

$$\mathrm{BIC}_k = -2L^* + k \log N \tag{2.10}$$

where L^* is the optimum likelihood value for a set including k haplotypes and N is the total number of observations in the dataset. Optimisation of the haplotype set was conducted for increasing values of k until the best set according to BIC was identified. We refer to this as the MLHapRec approach.

2.3.3 Validation of the Haplotype Reconstruction Method

2.3.3.1 Generation of Simulated Data

Simulated data were generated using a model of influenza transmission. Viruses were generated to have eight independent segments, of lengths equal to the segments of the A/H1N1 influenza virus. Each segment had five uniformly distributed polymorphic loci for which alternative and reference alleles were assigned at random. Through a process of generating all possible combinations of variant alleles, a total of 32 full length haplotypes were identified. Six haplotypes were randomly chosen from this set under the constraint that each of the five loci had to remain polymorphic, this being ensured through repeated sampling of all six haplotypes until the criteria was met. The frequencies of these haplotypes

were then randomly generated under the constraint of a minimum haplotype frequency of 5%, which was guaranteed through repeated sampling.

Each transmission event was modelled as a simple multinomial draw, selecting a number of viruses equal to the bottleneck size from the donor population. Identifying the new population frequencies as the rescaled transmission sample, within-host growth was then modelled as a second multinomial draw, conferring a 22-fold increase in the population size. Partial haplotype data were generated from simulated short reads of each viral segment. Short reads with lengths derived from a recent influenza dataset (Wilker et al. 2013) were generated (mean read length = 119.68, SD read length = 136.88, mean gap length = 61.96, SD gap length = 104.48, total read depth = 102825), these reads being used to calculate the number of reads spanning each set of consecutive polymorphisms in each segment. Given these numbers, full haplotype observations were generated using a Dirichlet-multinomial sampling process in the post-transmission population frequencies. Finally, partial haplotype observations were calculated by associating full haplotype draws with distinct partial haplotypes.

2.3.3.2 Results of Testing Against Simulated Data

Tested against simulated data, our method gave a reasonably good level of performance. Given data from simulated transmission events in which the population was comprised of 6 five-locus haplotypes, our method generally produced a reconstruction with the same number of haplotypes as this, reconstructed populations having a median of six haplotypes (Figure 2.12). Of the haplotypes that were inferred to exist, between 85% and 90% were found in the real population at some non-zero frequency. However, our method was not perfect, with some under- and over-calling of haplotypes, and imperfect identification of the true haplotypes.

2.3.4 Comparison of Haplotype Reconstruction Methods

A comparison of our two methods of haplotype reconstruction showed substantial differences in the number of haplotypes inferred to exist given increasingly sparse sequence reads from a population. Here we assess sparsity in terms of the number of variant alleles covered by a single read, defined by the mean read length from sequencing.

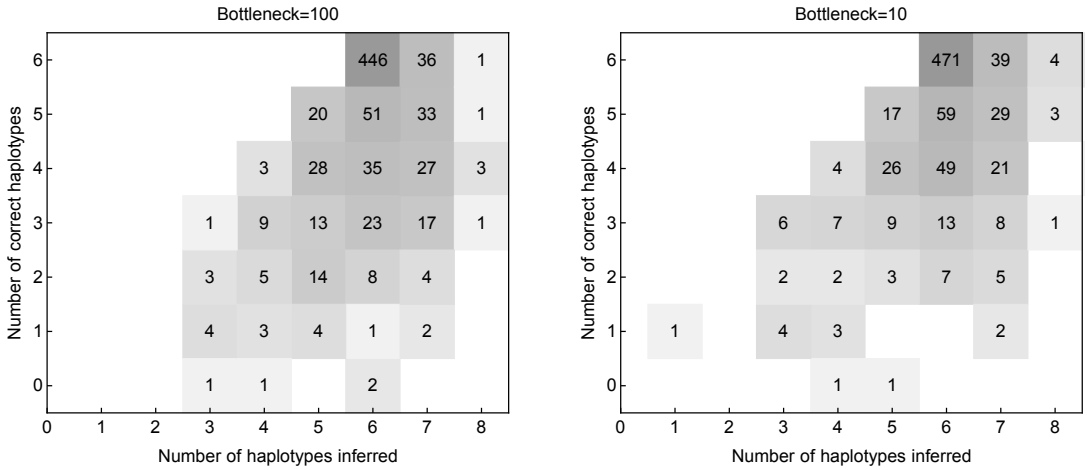


Figure 2.12. Numbers of inferred and correctly inferred haplotypes given simulated sequence data. A total of six haplotypes were included in each of 800 simulations tested.

To evaluate this statistic, sequence reads describing the gp41 region of a within-host HIV population, and with an original sequence length of 324 nucleotides were downloaded from a recent publication (Raghwani et al. 2018). Next, the set of reads was downsampled, removing short blocks of nucleotides of a given length from individual reads until no sufficiently long stretches of nucleotides remained; this process simulates the sequencing of the population with reads of length shorter than the original length. Finally, reads were used to reconstruct haplotypes using the SAMFIRE algorithm applied by Lumby et al. (2018) and using the new haplotype reconstruction method. Where longer reads were used, the number of haplotypes inferred was very similar, for example with 12 and 15 haplotypes respectively inferred given a read length of 300. However, at very short read lengths, SAMFIRE produced of the order 10^5 haplotypes whereas our new approach explained the available data with only three haplotypes (Figure 2.13). We note that the generation of large numbers of haplotypes is a deliberate feature of the SAMFIRE approach, whereas our minimal reconstruction method aims to generate a minimal number of haplotypes. Under the approach outlined in Figure 2.1, where n biallelic loci exist in a population, and reads describe only single loci, the code generates all 2^n possible haplotypes. While there are many advantages to this approach, this analysis shows that, in at least some cases, an approach which infers fewer haplotypes is required for calculations of the composition of a population to become feasible.

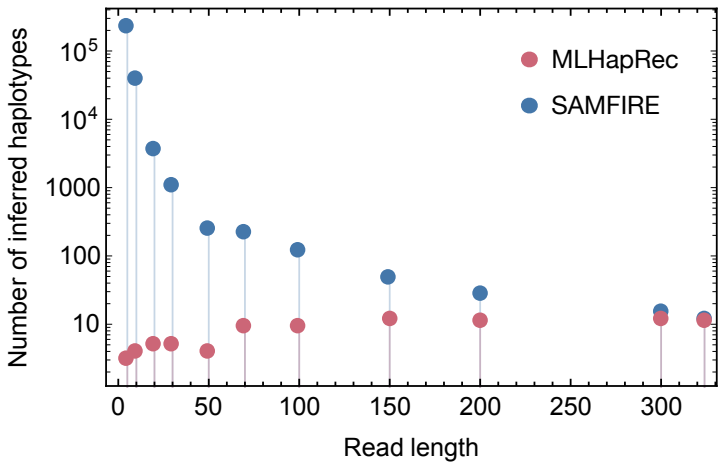


Figure 2.13. Number of reconstructed haplotypes obtained by the MLHapRec and SAMFIRE methods based on short-read sequence data from an HIV virus. An original dataset containing reads of length 324 bases was split to produce reads of shorter lengths before using each of the two reconstruction methods to produce an inferred set of haplotypes.

2.3.5 Comparison of Allele-Based and Haplotype-Based Transmission Inference Methods

Previous population genetic approaches to the inference of transmission bottlenecks have exploited an approach based upon changes in allele frequencies across transmission (Poon et al. 2016; Sobel Leonard et al. 2017a). We first explored a toy model of transmission in which bottleneck sizes were inferred using either allele-based or haplotype-based methods. In our toy model, a population was defined as having eight viral segments, each with two haplotypes. Before transmission, the frequency of the first haplotype on each segment i was set to $p_i = 0.5$. The population was assumed to be large, and well-mixed, with complete reassortment between segments implying a lack of linkage disequilibrium between alleles on different segments. Transmission was modelled as a binomial sampling process with a true bottleneck of $N^T = 100$, in which the new haplotype frequency p_i' was defined by the probability of n_i copies of the first haplotype of segment i being transmitted. We assumed that the new haplotype frequencies were observed in a noise-free manner:

$$P\left(p_i' = \frac{n_i}{N^T}\right) = \frac{N^T!}{n_i!(N^T - n_i)!}\left(\frac{1}{2}\right)^{N^T} \tag{2.11}$$

A haplotype-based inference of the bottleneck size was conducted using a simple maximum likelihood calculation, approximating the likelihood as a nor-

mal distribution with mean and variance equal to the binomial sampling process:

$$\log L(N|p_i') = \log \mathcal{N}\left(p_i' \middle| 0.5, \sqrt{\frac{p_i(1-p_i)}{N}}\right) \tag{2.12}$$

where

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.13}$$

In our haplotype-based inference, the value of N was inferred by identifying the maximum likelihood value of

$$L_{\boldsymbol{h}} = \sum_{i=1}^{8} \log L(N|p_i') \tag{2.14}$$

where the sum is over the eight viral gene segments.

The allele-based bottleneck inference worked in a similar manner to this approach. Via simulation we construct a system whereby the first seven viral segments contain a single polymorphism, but the final segment contains ten polymorphisms (Figure 2.14A). Under an allele-based approach to the data, these polymorphisms are assumed to be independent. On the final segment, we note that, as only two haplotypes exist, the linked alleles at all of the polymorphic loci will be observed at identical frequencies. The dataset on which this inference is performed is therefore identical to that used for the haplotype-based method, albeit that the final haplotype frequency p_8' is observed ten times, once for each polymorphism. This results in the allele-based likelihood

$$L_A = \sum_{i=1}^{7} \log L(N|p_i') + 10 \log L(N|p_8') \tag{2.15}$$

Simulated data were generated for 5000 transmission events and the optimal population bottleneck identified in each case by evaluating likelihoods for bottlenecks in the range of $[1, 1000]$.

2.3.5.1 Results From the Toy Model

The results of a toy model showed that under idealised circumstances a haplotype-based model of bottleneck inference outperforms an equivalent allele-based model. In our evaluation, both the allele-based and the haplotype-based methods inferred a value for the bottleneck that was correct in the mean; across 5000 inferences the harmonic mean values of the inferred bottlenecks were 100.7 and 100.2

respectively. However, the haplotype-based method gave more precise inferred bottlenecks, with a 95% range in the inferred values between 46 and 337, compared to a range between 30 and 552 for the allele-based method (Figure 2.14B). This result can be simply understood. Transmission is a stochastic process, in which the frequency of the haplotypes in each segment changes between the donor and the recipient hosts. Both of our inference methods evaluate changes in these frequencies to infer the size of the bottleneck that occurred, using a maximum likelihood approach. Combining the information across segments gives a collective insight into the bottleneck. Where the population is considered as haplotypes, this is achieved correctly; each segment has an equal weight in discerning what is the true bottleneck. However, where the population is considered as a set of independent alleles, this does not occur; the final segment is falsely given a higher weighting in the likelihood calculation than the others. The change in the frequency of haplotypes in the final segment thus dominates the likelihood calculation. Changes in a single segment are more prone to stochasticity than the combined changes across multiple segments. The allele-based method therefore produces inferences which have a lower degree of precision.

Estimates of the uncertainty in an inference may differ substantially between the allele-based and the haplotype-based methods. In falsely treating allele frequencies as being independent of one another, the allele-based method produces a log likelihood that is larger in magnitude than the correct likelihood for the system (compare Equations 2.14 and 2.15); this can lead to an over-optimistic assessment of the confidence with which the size of the bottleneck can be estimated. Given a bottleneck size of 100 transmitted virions and initial haplotype frequencies of 50% for each segment in Figure 2.14A, the expected change in allele frequency across the transmission event is 5%. Applying the allele- and haplotype-based methods to simulated data describing such an event gave correct inferences of the bottleneck in each case, but with dramatically different confidence intervals; a cutoff of two log likelihood units in the allele-based method gave the interval (71,136), but an interval of (30,236) was produced by the haplotype-based model (Figure 2.15).

Having highlighted the benefit of haplotype-based approaches to bottleneck inference, we now evaluate our new approach to this task.

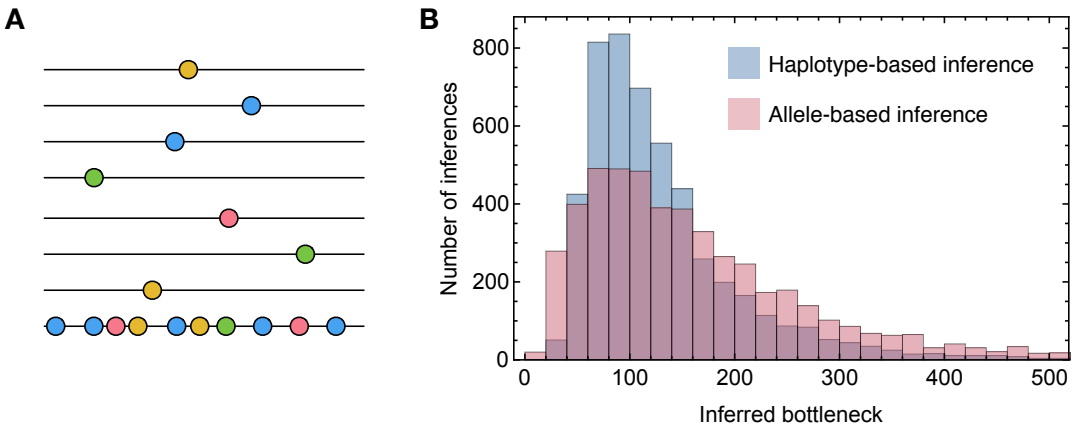


Figure 2.14. **A** Illustration of our toy model system used in transmission. A virus contains eight segments. The population of each segment comprises two haplotypes, which differ from one another at a number of loci; in seven segments the haplotypes differ by a single polymorphism, while in the final segment the haplotypes differ at ten polymorphisms. **B** Inferred bottlenecks from simulations of the transmission of our toy model system, calculated using either a haplotype-based or allele-based method of inference. The haplotype-based model gives a more precise inference of the transmission bottleneck.

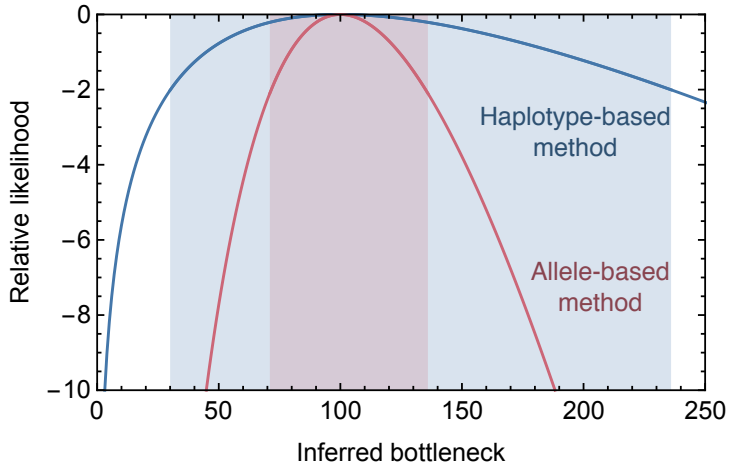


Figure 2.15. Likelihood functions from haplotype-based and allele-based inference methods for a case in which our toy transmission model describes a change in haplotype frequency equal to the expectation for each segment for a bottleneck of size 100. While each method correctly diagnoses the transmission bottleneck, the allele-based method produces an overly-optimistic estimate of the confidence interval with which the bottleneck can be inferred.

2.3.6 Full Transmission Model: Methods

A qualitative illustration of our transmission model is shown in Figure 2.16. The population before transmission is represented by the hidden state $\boldsymbol{q}^B$; this population transmits to form the founder population $\boldsymbol{q}^F$ which grows to the after-

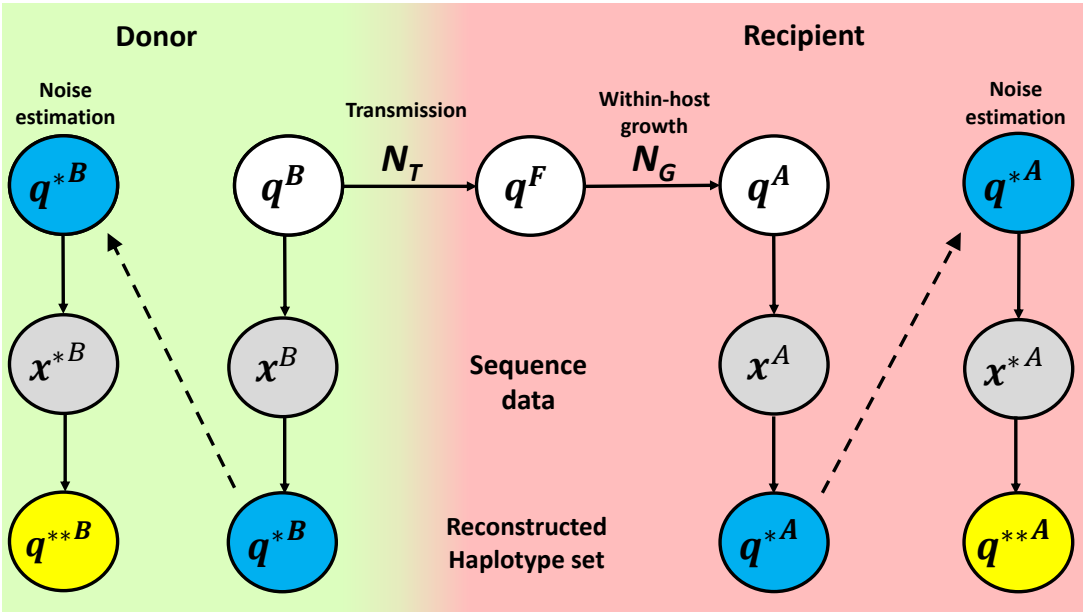


Figure 2.16. Overview of the transmission model.

transmission viral population $\boldsymbol{q}^A$. These respective steps are represented by the transmission bottleneck N^T and the within-host growth effective population size N^G. Populations are observed before and after transmission to give the datasets $\boldsymbol{x}^B$ and $\boldsymbol{x}^A$, from which we build the haplotype reconstruction populations $\boldsymbol{q}^{*B}$ and $\boldsymbol{q}^{*A}$; these give the mean values of our estimates of $\boldsymbol{q}^B$ and $\boldsymbol{q}^A$.

We next assess the extent to which noise in the observation affects our haplotype reconstructions. Taking each of $\boldsymbol{q}^{*B}$ and $\boldsymbol{q}^{*A}$, we artificially simulate 100 datasets $\boldsymbol{x}_i^{*B}$ and $\boldsymbol{x}_i^{*A}$, each having the same properties as $\boldsymbol{x}^B$ and $\boldsymbol{x}^A$, and being generated as observations from $\boldsymbol{q}^{*B}$ and $\boldsymbol{q}^{*A}$. From each of these datasets we perform haplotype reconstruction under the assumption that the underlying haplotypes are correct (that is, simply learning their frequencies), generating 100 inferences $\boldsymbol{q}_i^{**B}$ and $\boldsymbol{q}_i^{**A}$. Comparing $\boldsymbol{q}_i^{**B}$ to $\boldsymbol{q}_i^{*B}$ and $\boldsymbol{q}_i^{**A}$ to $\boldsymbol{q}_i^{*A}$, we derive variances in $\boldsymbol{q}^{*B}$ and $\boldsymbol{q}^{*A}$, which for simplicity we assume to be represented by diagonal covariance matrices. As $\boldsymbol{x}_i^{*B}$ and $\boldsymbol{x}_i^{*A}$ were emitted from $\boldsymbol{q}_i^{*B}$ and $\boldsymbol{q}_i^{*A}$ in a manner identical to the way $\boldsymbol{x}_i^B$ and $\boldsymbol{x}_i^A$ were emitted from $\boldsymbol{q}_i^B$ and $\boldsymbol{q}_i^A$, we assume that the variances in $\boldsymbol{q}_i^B$ and $\boldsymbol{q}_i^A$ equal the variances in $\boldsymbol{q}_i^{*B}$ and $\boldsymbol{q}_i^{*A}$. Having thus identified mean and variances for $\boldsymbol{q}^B$ and $\boldsymbol{q}^A$ we employ these for the calculation of transmission bottlenecks.

Using our reconstructed haplotype set, $\boldsymbol{h}$, to calculate the likelihood of the bottleneck size given data, assuming a completely neutral transmission, we take

a similar approach to that of Lumby et al. (2018) (to be explained further in the next chapter) by splitting the likelihood into two components:

$$\begin{aligned} L(N^T|\boldsymbol{q}^{*B},\boldsymbol{q}^{*A},N^G) = &\int P(\boldsymbol{q}^{*B}|\boldsymbol{q}^B)P(\boldsymbol{q}^B)d\boldsymbol{q}^B \\ &\times \int P(\boldsymbol{q}^{*A}|\boldsymbol{q}^A)\Bigg\{\int P(\boldsymbol{q}^A|N^G,\boldsymbol{q}^F) \\ &\times \left(\int P(\boldsymbol{q}^F|N^T,\boldsymbol{q}^B)P(\boldsymbol{q}^B)d\boldsymbol{q}^B\right)d\boldsymbol{q}^F\Bigg\}d\boldsymbol{q}^A \end{aligned} \tag{2.16}$$

where the first integral corresponds to the initial observation of the system and the subsequent ones encompass transmission, within-host growth and post-transmission sampling.

Interpreting the quantities in Figure 2.16 as random variables, and noting that transmission and sampling processes can be represented by multinomial and Dirichlet-multinomial outcomes respectively, we approximate these discrete distributions as continuous multivariate normal distributions and evaluate the integrals in Equation 2.16 through the means of compound distributions. Given distributions F in x and G in y, a compound distribution H takes the form

$$P_H(x) = \int P_F(x|y)P_G(y)\mathrm{d}y \tag{2.17}$$

where the mean and variance of H are defined by the law of total expectation,

$$\mathrm{E}_H[x] = \mathrm{E}_G[\mathrm{E}_F[x|y]], \tag{2.18}$$

and the law of total variance,

$$\mathrm{var}_H[x] = \mathrm{E}_G[\mathrm{var}_F[x|y]] + \mathrm{var}_G[\mathrm{E}_F[x|y]], \tag{2.19}$$

respectively.

As mentioned above, for the pre-transmission component we identify

$$\begin{aligned} \mathrm{E}[q_i^B] = \mathrm{E}[q_i^{*B}] = \mu_i^B, \\ \mathrm{var}[q_i^B] = \mathrm{var}[q_i^{*B}] = \left(\sigma_i^B\right)^2 \end{aligned} \tag{2.20}$$

where the subscript i denotes haplotype i. The mean μ_i^B is evaluated by optimising the frequency of the reconstructed haplotype set using the MLHapRec approach and $\left(\sigma_i^B\right)^2 = \sum_{j=1}^{100}(q_i^{*B} - q_j^{**B})^2/100$ is the variance in the frequency of q_i^{*B} calculated numerically over $r = 100$ replicate samples. We assumed that all the off-diagonal elements of the covariance matrix are zero, which is equivalent to disregarding between-haplotype correlations in specifying the uncertainty in μ_i^B.

Moving on to the post-transmission component of the compound distribution in Equation 2.16, we can carry out the relevant marginalisations using the law of total expectation and the law of total variance.

Given that the dynamics governing transmission and within-host growth are assumed selectively neutral, the mean frequencies of the viral population are unchanged following transmission and growth. Specifically, the transmission event can be modelled as a single multinomial draw with N^T number of trials. As a result, the conditional distribution for the founder population can be represented by a multivariate normal with mean

$$\mathrm{E}[q_i^F|q_i^B] = q_i^B, \tag{2.21}$$

and variance

$$\mathrm{var}[q_i^F|q_i^B] = \frac{q_i^B(1-q_i^B)}{N^T}. \tag{2.22}$$

assuming that the variance can be represented by a diagonal matrix. Therefore, marginalising over q_i^B yields a mean

$$\mathrm{E}[q_i^F] = \mathrm{E}[\mathrm{E}[q_i^F|q_i^B]] = \mathrm{E}[q_i^B] = \mu_i^B, \tag{2.23}$$

and variance of

$$\begin{aligned}\text{var}[q_i^F] &= \text{E}[\text{var}[q_i^F|q_i^B]] + \text{var}[\text{E}[q_i^F|q_i^B]] \\ &= \text{E}[\frac{q_i^B(1-q_i^B)}{N^T}] + \text{var}[q_i^B] \\ &= \frac{1}{N^T}\mu_i^B(1-\mu_i^B) + \left(1-\frac{1}{N^T}\right)\left(\sigma_i^B\right)^2. \end{aligned} \tag{2.24}$$

The within-host growth dynamics can be modelled as a multinomial draw of depth $N^G = gN^T$ where g is the growth factor. The concept of the growth factor will be discussed in detail in Chapters 3 and 4. Fixing the growth factor as $g = 22$ (Lumby, Nene and Illingworth 2018), we define the conditional mean of q_i^A as

$$\text{E}[q_i^A|q_i^F] = q_i^F, \tag{2.25}$$

and the variance as

$$\text{var}[q_i^A|q_i^F] = \frac{q_i^F(1-q_i^F)}{N^G}. \tag{2.26}$$

We can subsequently marginalise over q_i^F to find the mean

$$\text{E}[q_i^A] = \text{E}[\text{E}[q_i^A|q_i^F]] = \text{E}[q_i^F] = \mu_i^B \tag{2.27}$$

and variance

$$\begin{aligned}\text{var}[q_i^A] &= \text{E}[\text{var}[q_i^A|q_i^F]] + \text{var}[\text{E}[q_i^A|q_i^F]] \\ &= \text{E}[\frac{q_i^F(1-q_i^F)}{N^G}] + \text{var}[q_i^F] \\ &= \frac{1}{N^G}\mu_i^B(1-\mu_i^B) + \left(1-\frac{1}{N^G}\right)\Big\{\frac{1}{N^T}\mu_i^B(1-\mu_i^B) + \left(1-\frac{1}{N^T}\right)\left(\sigma_i^B\right)^2\Big\} \end{aligned} \tag{2.28}$$

for the resulting moments of q_i^A.

As discussed above, we define the conditional moments of q_i^{*A} as

$$\mathrm{E}[q_i^{*A}|q_i^A] = q_i^A, \tag{2.29}$$

for the mean and

$$\mathrm{var}[q_i^{*A}|q_i^A] = \left(\sigma_i^A\right)^2, \tag{2.30}$$

for the variance. Finally we may then marginalise over q_i^A to find the resulting mean and variance for the post-transmission component in Equation 2.16. For the mean we get

$$\mathrm{E}[q_i^{*A}] = \mathrm{E}[\mathrm{E}[q_i^{*A}|q_i^A]] = \mathrm{E}[q_i^A] = \mu_i^B, \tag{2.31}$$

whilst the variance expression yields

$$\begin{aligned}\mathrm{var}[q_i^{*A}] &= \mathrm{E}[\mathrm{var}[q_i^{*A}|q_i^A]] + \mathrm{var}[\mathrm{E}[q_i^{*A}|q_i^A]] = \mathrm{E}\left[\left(\sigma_i^A\right)^2\right] + \mathrm{var}[q_i^A] \\ &= \left(\sigma_i^A\right)^2 + \frac{1}{N^G}\mu_i^B(1-\mu_i^B) + \left(1-\frac{1}{N^G}\right)\Big\{\frac{1}{N^T}\mu_i^B(1-\mu_i^B) + \\ &\qquad \left(1-\frac{1}{N^T}\right)\left(\sigma_i^B\right)^2\Big\}\end{aligned} \tag{2.32}$$

where $\left(\sigma_i^A\right)^2$ is the variance obtained from q_i^{*A}. This variance is defined in a manner similar to that of $\left(\sigma_i^B\right)^2$ and may be found numerically as $\left(\sigma_i^A\right)^2 = \sum_{j=1}^{100}(q_i^{*A} - q_j^{**A})^2/100$.

Together, Equations 2.31 and 2.32 define the mean and variance of a multivariate normal distribution representing the post-transmission component of the likelihood in Equation 2.16. Given our inferences for q_i^{*B} and q_i^{*A}, we optimised the likelihood with respect to N^T, considering bottlenecks in the range $N^T \in [1, 1000]$, producing a maximum likelihood estimate for the size of the transmitted population.

2.3.6.1 Results From Experimental Data

Our transmission model was applied to data collected from a household study previously published by McCrone et al. (2018). This study previously identi-

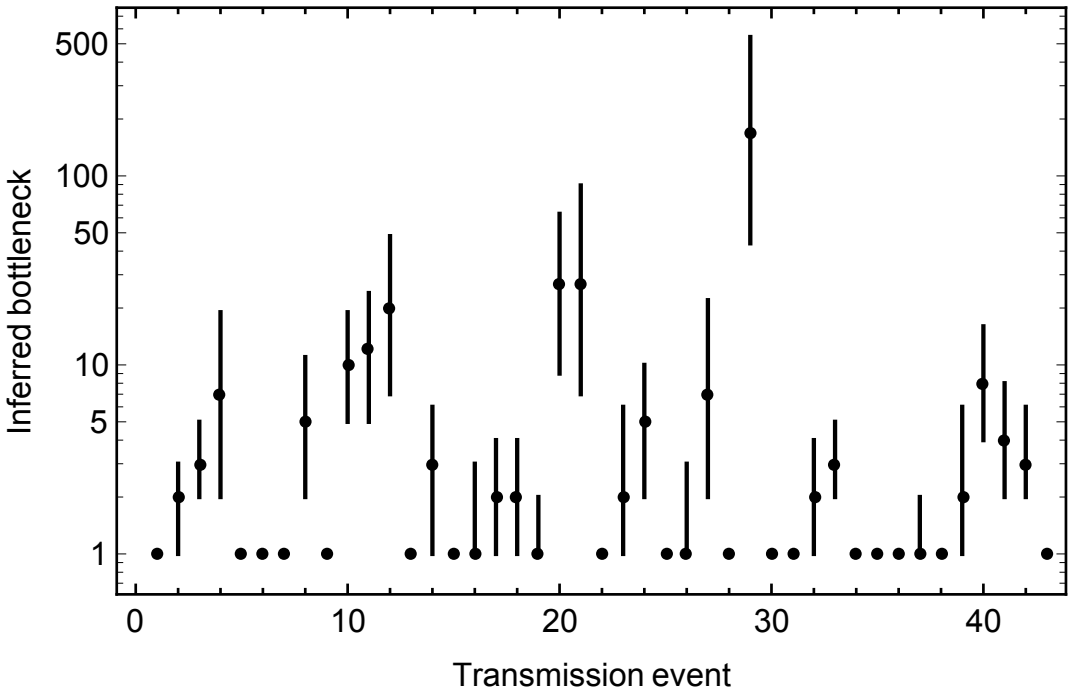


Figure 2.17. Bottleneck sizes inferred from the data presented by McCrone et al. Dots indicate the maximum likelihood bottleneck size inferred for each of the 43 systems described in this work. Vertical bars represent confidence intervals equivalent to a cut-off of 2 log likelihood units.

fied narrow bottlenecks in human-to-human transmission, with all but a single transmission being inferred to involve between one and four viral particles. Application of our haplotype-based method generated a broader range of inferred bottleneck values (Figure 2.17), albeit one that reflected the general pattern of results from the first analysis. In our results, the most common inference was that infection was initiated by a single viral particle, this being inferred in 21 of the 43 transmission events studied. Furthermore, a collective analysis, in which a single bottleneck was fitted to all of the data as a whole, produced an inferred bottleneck of $N^T = 1$. Our results differed from the original analysis in that generally higher bottlenecks were inferred, with six events being inferred to involve a bottleneck of 10 particles or more. Our inference supports a general rule for influenza transmission in which the majority of transmission events involve a founder population that comprises a very small number of viral particles, with a few exceptions occurring, where the founding population is larger than this.

2.4 Discussion

I have here outlined a framework for the inference of haplotype frequencies from data, based upon a likelihood scheme incorporating partial haplotype information, and demonstrated various applications of this approach to evolutionary inference. Our initial results demonstrated the potential to make haplotype inferences in cases where a small number of haplotypes were present in a system. Applied to within-host data describing HIV adaptation we showed that the redundancies inherent to our haplotype reconstruction did not have a large impact on the energy scores derived from a global measure of viral fitness; this indicates the potential for such measures, applied to within-host evolution, to be employed to generate insight into viral evolution. Our second application, to an influenza B infection in an immunosuppressed host, highlighted the potential for our methods to gain insight into populations going beyond what can be achieved via consensus sequence-based methods. Here we showed that populations that were phylogenetically distinct with regards to their consensus sequence were also distinct when considered at the level of variants within the viral population. Finally, we outlined an approach for identifying minimal haplotype reconstructions of data when applied to inferences of viral transmission; we showed the inherent advantage of such an approach when applied to sparse sequence data, and demonstrated that haplotype-based methods, by making a proper account of the fact that viruses, rather than independent alleles, are transmitted, grant an improved inference of transmission bottleneck size. Applying our method to published data describing influenza transmission in a household study, we achieved slightly different results from those previously published, but which nevertheless support a general pattern of small numbers of viral particles generally founding new infections. Haplotype reconstruction therefore provides a broadly useful tool generating biological insight into a number of systems.

Having demonstrated the potential of haplotype reconstruction methods, I now turn in more detail to the question of viral transmission, which will underpin the remainder of this thesis. I note that our inference of transmission bottlenecks neglects the impacts of selection upon transmission. During a transmission event, both selection and genetic drift may cause changes in the diversity of a viral population. In the next chapter I outline a method to separate these changes, inferring bottleneck sizes in the presence of natural selection.

Chapter 3

Basic Transmission Inference Scheme and Application to Simulated Data

3.1 Introduction

The previous chapter considered the reconstruction of haplotypes and the inference of haplotype frequencies from short-read data. In this chapter I introduce the basic transmission inference scheme upon which this and subsequent chapters rely. The transmission model considers changes in genetic diversity as a means of inferring transmission events with the viral population represented in terms of viral haplotypes and their associated frequencies. Existing approaches for transmission inference tend to neglect the impact of selection upon the viral population. I here highlight a need for future methods to acknowledge the importance of both selection and bottleneck effects upon viral diversity and to define techniques for distinguishing these. In this chapter I present our solution to this problem and investigate the performance of our algorithm when applied to simulated data.

A handful of methods for transmission inference have been developed (Khiabanian et al. 2015; Krimbas and Tsakas 1971; Monsion et al. 2008; Sacristan et al. 2003), many of which are based upon the analysis of diversity changes from single-locus variant data. These methods assume independence between variant sites, which, as we have seen in the previous chapter, leads to potential biases. I here compare the performance of our basic transmission scheme to the

beta-binomial method of Sobel Leonard et al. (2017b), which, to our knowledge, represents the present state-of-the-art for bottleneck inference.

3.1.1 Transmission Inference

Understanding viral transmission is a key task for viral epidemiology. The extent to which a virus is able to transmit between hosts determines whether it is likely to cause sporadic, local outbreaks, or spread to cause a global pandemic (Breban, Riou and Fontanet 2013; Fraser et al. 2009). In a transmission event, the transmission bottleneck, which specifies the number of viral particles founding a new infection, influences the amount of genetic diversity that is retained upon transmission, with important consequences for the evolutionary dynamics of the virus (Bergstrom, McElhany and Real 1999; Gutiérrez, Michalakis and Blanc 2012).

Recent studies have used genome sequencing approaches to study transmission bottlenecks in influenza populations. In small animal studies, the use of neutral genetic markers has shown that the transmission bottleneck is dependent upon the route of transmission, whether by contact or aerosol transmission (Frise et al. 2016; Varble et al. 2014). In natural human influenza populations, where modification of the virus is not possible, population genetic methods have been used to analyse bottleneck sizes. Analyses of transmission have employed different approaches, exploiting the observation or non-observation of variant alleles (Sacristan et al. 2003) or using changes in allele frequencies to characterise the bottleneck under a model of genetic drift (Charlesworth 2009; Khiabanian et al. 2015; Krimbas and Tsakas 1971; Monsion et al. 2008). A recent publication improved this latter model, incorporating the uncertainty imposed upon allele frequencies by the process of within-host growth (Sobel Leonard et al. 2017b). Two studies of within-household influenza transmission have provided strikingly different outcomes in the number of viruses involved in transmission, with estimates of 1-2 (McCrone et al. 2018) and 100-200 (Poon et al. 2016) respectively, albeit that the data used to generate the latter result has recently been challenged (Xue and Bloom 2018).

Another focus of research has been the role of selection during a transmission event; this is important in the context of the potential for new influenza strains to become transmissible between mammalian hosts (Kuiken et al. 2006; Lipsitch et al. 2016). Studies examining transmissibility have assessed the potential for different strains of influenza to achieve droplet transmission between

ferrets under laboratory conditions (Herfst et al. 2012; Imai et al. 2012; Sutton et al. 2014; Yang et al. 2016); ferrets provide a useful, if imperfect, model for transmission between humans (Buhnerkempe et al. 2015; Palese and T. T. Wang 2012). The application of bioinformatic techniques to data from these experiments has identified 'selective bottlenecks' in the experimental evolution of these viruses (Moncla et al. 2016; Wilker et al. 2013), whereby some genetic variants appear to be more transmissible than others. In these studies, selection has been considered in terms of the population diversity statistic π; changes in π_N/π_S, the ratio between non-synonymous and synonymous diversity, have been used to evaluate patterns of selection across different viral segments.

I here note the need for a greater clarity of thinking in the analysis of viral transmission events. For example, analysis of genetic variants in viral populations shows that synonymous and non-synonymous mutations both have fitness consequences for viruses (Acevedo, Brodsky and Andino 2014; Visher et al. 2016); the use of synonymous variants as a neutral reference set may not hold. More fundamentally, in an event where the effective population size is small, the influences of selection and genetic drift may be of similar magnitude (Rouzine, Rodrigo and Coffin 2001). However selection is assessed, this implies a need to separate stochastic changes in a population from selection, especially where a transmission bottleneck may include only a small number of viruses (McCrone et al. 2018; Varble et al. 2014; Zwart, Daròs and Elena 2011). It is possible for the attribution of a change in diversity to the action of selection, or the attribution of allele frequency change to genetic drift to be flawed. Given the increasing availability of sequence data, more sophisticated tools for the analysis of viral transmission are required.

Here I observe three challenges in the analysis of data from viral transmission events. Firstly, selection can produce changes in a population equivalent to those arising through a neutral population bottleneck (Abel et al. 2015) (Figure 3.1A), making it necessary to distinguish between the two scenarios. A broad literature has considered the simultaneous inference of the magnitude of selection acting upon a variant along with an effective population size (Bollback, York and Nielsen 2008; Feder, Kryazhimskiy and Plotkin 2014; Foll et al. 2014; Malaspinas et al. 2012; O'Hara 2005; Terhorst, Schlötterer and Song 2015). However, such approaches rely on the observation of an allele frequency at more than two time points so as to distinguish a deterministic model of selection (with an implied infinite effective population size) from a combined model of selection

with genetic drift; such approaches cannot be directly applied to the analysis of viral transmission.

Secondly, inferences of transmission events need to account for the haplotype structure of viral populations, whereby whole viruses, rather than sets of independent alleles, are transmitted (Figure 3.1B). The low rate of homologous recombination in segments of the influenza virus (Boni et al. 2008; Chare, Gould and Holmes 2003) implies that viral evolution proceeds at the haplotype level (Neher and Shraiman 2009); competition occurs between collections of linked alleles, or segments, rather than the individual alleles themselves. Under such circumstances, fitter variants do not always increase in frequency within a population (Illingworth and Mustonen 2012b; Koelle and Rasmussen 2015; Strelkowa and Lässig 2012). Calculations of genetic drift, which are often derived from the evolution of independent variants (Felsenstein 1971), need to be adjusted to account for this more complex dynamics. While haplotype reconstruction as a basis for bottleneck inference was described in the previous chapter, the need to account for selection leads to the use of a more generous haplotype reconstruction approach, allowing for the existence of haplotypes into which the population can move under the influence of selection.

Thirdly, noise in the measurement of a population may influence the inferred size of a transmission bottleneck (Figure 3.1C). A broad range of studies have examined the effect of noise in variant calling and genome sequence analysis (Beerenwinkel and Zagordi 2011; Dijk et al. 2014; Iyer et al. 2015; Laehnemann, Borkhardt and McHardy 2016; McCrone and Lauring 2016; Sandmann et al. 2017; Varghese et al. 2010; C. Wang et al. 2007); more recently formulae have been proposed to measure the precision with which allele frequencies can be defined given samples from a population (Illingworth 2015; Illingworth et al. 2017; Zanini et al. 2017). Where small changes in allele frequencies are used to assess a population bottleneck, it is important to separate the effects of noise in the measurement of populations from genuine changes in a population.

In this chapter I describe a novel method for the inference of population bottlenecks in influenza which addresses the above issues. I show that our approach correctly evaluates changes in a population even where the data describing that change are affected by noise. In common with the approach of the previous chapter, it explicitly accounts for the haplotype structure of a population, using a method of haplotype reconstruction as a framework within which to evaluate data present within short sequence reads. However, in addition, where these factors can be discriminated, the method presented here distinguishes

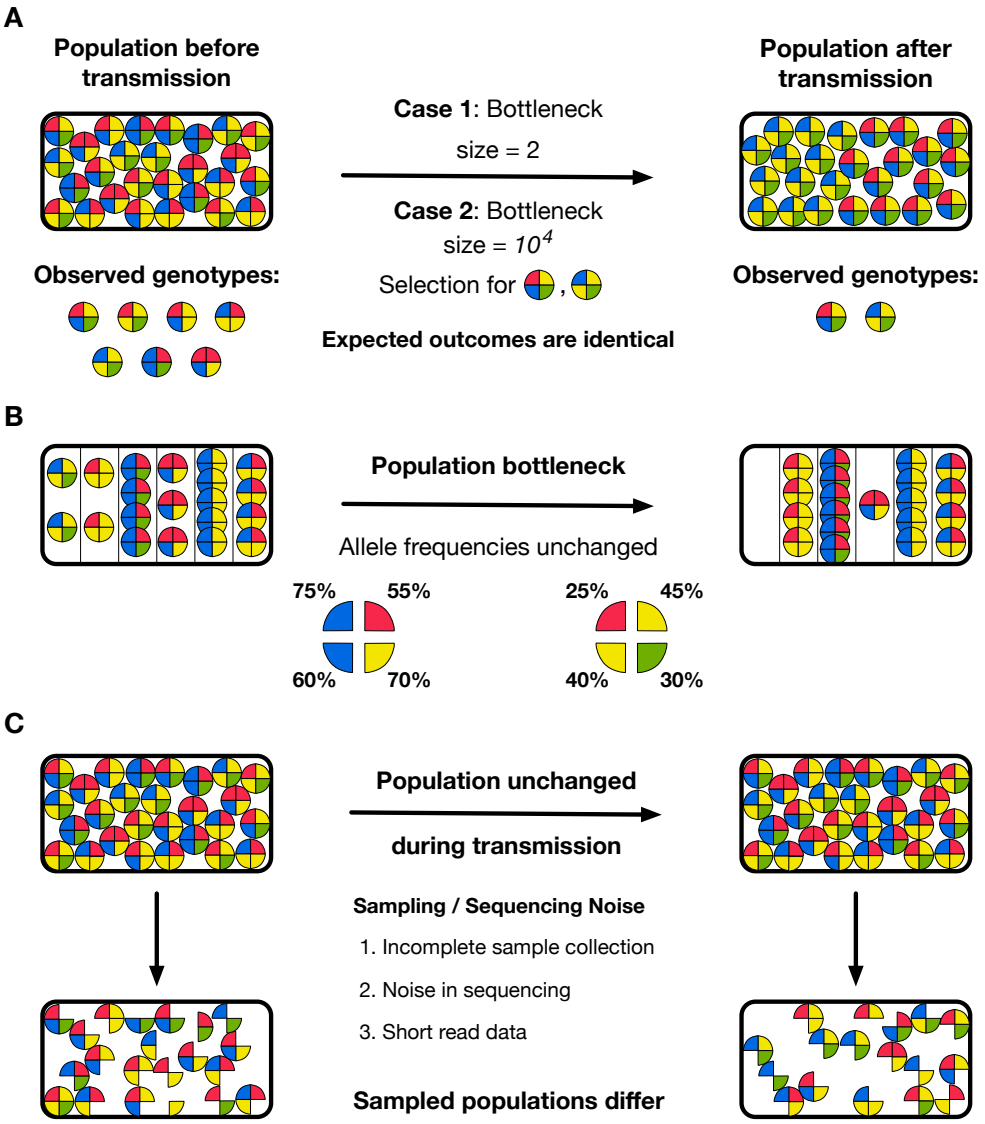


Figure 3.1. Challenges arising in the inference of transmission bottlenecks from viral sequence data. Circles represent idealised viral particles characterised by four distinct alleles. **A.** Reductions in population diversity cannot necessarily be attributed unambiguously to either a population bottleneck, or the action of selection. In the illustrated case, either a tight bottleneck without selection or a large bottleneck with strong selection could explain the change in the population during transmission. **B.** Straightforward statistics describing a population may generate misleading inferences of population bottleneck size. In the illustrated case, the genetic structure of a population is changed by a population bottleneck during transmission, but the frequency of each allele within the population does not change; an inference of bottleneck size derived from single-locus statistics would incorrectly be very large. **C.** Noise arising from the process of collecting and sequencing data is likely to produce differences between the observed populations, even in the event that the composition of the viral population was entirely unchanged during transmission.

between the influences on the population of selection and the transmission bottleneck. Studies of viral evolution have highlighted the potential for payoffs between within-host viral growth and transmissibility (Blanquart et al. 2016); given sufficient data we can evaluate how selection operates upon each of these phenotypes. Our model extends previous population genetic work on bottleneck inference to provide a more generalised model for the analysis of data spanning viral transmission events.

3.1.2 Author Contributions

This chapter is adapted from material which formed the basis of a published paper (Lumby, Nene and Illingworth 2018). Nuno Nene and Christopher Illingworth contributed to general discussions during model development but the great majority of the work was completed by the author.

3.2 Methods

3.2.1 Model Outline

In the recent literature, the term 'bottleneck' has been applied to describe a reduction in the genetic diversity of a population (e.g. (Zaraket et al. 2015)), whether arising from selection or a numerical reduction in the size of a population. Here, we define a 'bottleneck' more strictly as a neutral process whereby a finite number of viral particles from one population found a subsequent generation of the population, either within the same host, or across a transmission event from host to recipient. Selection then constitutes a modification to this process whereby some viruses, because of their genotype, have a higher or lower probability of making it through the bottleneck to found the next generation.

Building on the haplotype inference methods described in the previous chapter, I developed a population genetic method for making a joint inference of the bottleneck size and the extent of selection acting during a transmission event. In my basic model of transmission (Figure 3.2), I consider a setup wherein a viral population is transmitted from one host to another with samples being collected before and after transmission. Within this model, viruses are categorised as haplotypes according to the alleles they harbour at polymorphic sites in the genome. Haplotypes are constructed from short-read data via the exhaustive method (Illingworth 2015, 2016) outlined in Chapter 2. The viral population

is then represented as a vector of frequencies of haplotypes in this set; the population before transmission is represented by the vector $\boldsymbol{q}^B$ (B denoting 'Before transmission'). During transmission, a random sample of N^T viruses are passed on to the second host to give the founder population $\boldsymbol{q}^F$. Selection for transmissibility, whereby genetic variants cause some viruses to be more transmissible than others, is described by the function S^T. The potentially small size of the founder population means that the population evolves within the host under the influence of genetic drift to create the large post-transmission population $\boldsymbol{q}^A$ (A denoting 'After transmission'); this process is approximated in our model by a Wright-Fisher sampling process (representing genetic drift) with effective population size N^G. Observations of the population are collected before and after transmission via a noisy sequencing process to give the datasets $\boldsymbol{x}^B$ and $\boldsymbol{x}^A$. The extent of noise in the sampling and sequencing is characterised by the parameter C (Illingworth 2015, 2016). Noise in our study was considered in terms of the precision with which the frequency of a variant can be specified by viral sequence data. Variant frequencies are measured in terms of the number of reads which report a given allele; in the absence of noise the uncertainty in the frequency would be that arising from a binomial distribution. Our noise parameter C describes the extent to which this uncertainty is increased. Smaller values of C increase the variance, reaching that of a non-informative uniform distribution at $C = 0$ whilst larger values represent lesser additional uncertainty, tending towards the binomial limit as $C \to \infty$ (Figure 3.3). The parameter C and the absolute read depth of a sample can be converted into an 'effective depth' of sequencing, see (Illingworth et al. 2017). In the limit of very deep sequencing the variance of an allele frequency tends towards that of a binomial distribution with sampling depth $C + 1$.

3.2.2 Differentiating Selection From Drift

We note that both the transmission and within-host growth events can be represented as sampling processes, with the former potentially biased by the effect of selection. As such, given an estimate of the noise inherent to the sequencing process and externally-determined estimates for N^G, we can calculate an approximate likelihood for the parameters N^T and S^T given the observations $\boldsymbol{x}^B$ and $\boldsymbol{x}^A$. Maximising this likelihood gives an estimate for the size of the transmission bottleneck and the extent to which specific genetic variants within the pre-transmission population confer increased transmissibility upon viruses.

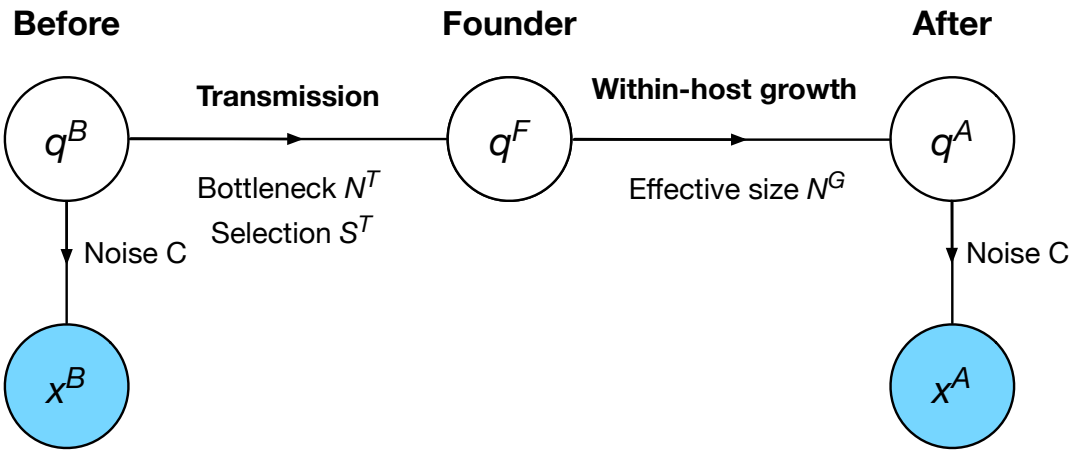


Figure 3.2. Basic model of transmission. A set of haplotypes exists at frequencies $\boldsymbol{q}^B$ from which a noisy observation $\boldsymbol{x}^B$ is made. During a transmission event, a total of N^T viruses are transferred under the influence of selection S^T, establishing an infection in the next host described by $\boldsymbol{q}^F$. Growth of the viral population within the host then occurs to produce the population $\boldsymbol{q}^A$, influenced by genetic drift (characterised by the effective population size N^G). Sampling of the final population gives the second observation $\boldsymbol{x}^A$.

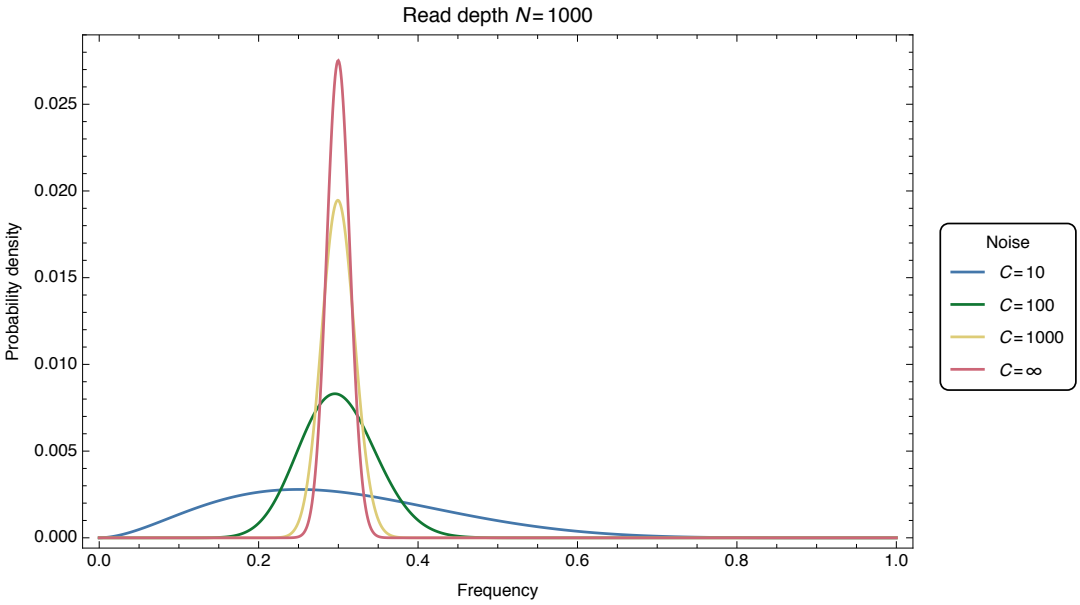


Figure 3.3. Allele frequency distribution for a sample of read depth $N = 1000$ collected from a population with true allele frequency one third, with a noise-free sampling method ($C = \infty$) and with C values of 10, 100, and 1000.

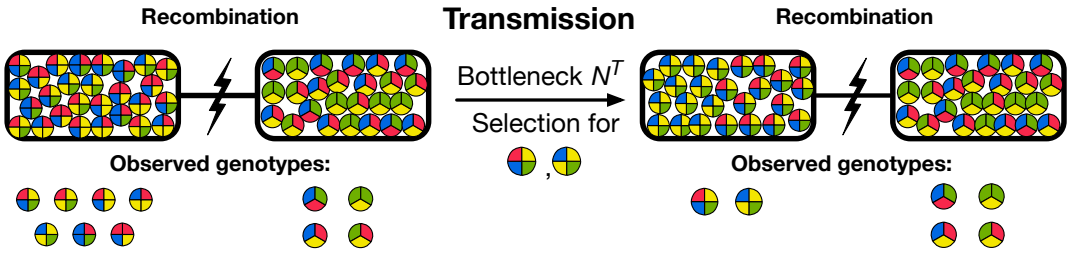


Figure 3.4. Regions of the genome which are separated by recombination or reassortment are used to distinguish the effects of selection and a population bottleneck. Prior to transmission, the first region contains seven different genotypes spanning four variant loci whilst the second region harbours four genotypes covering three loci. As recombination between these two regions leaves them unlinked, selection acting on genotypes in one region has no impact on the fate of genotypes in the other region. Thus, where genetic diversity is reduced in the first region, the preservation of diversity in the second region attributes this change to the action of selection on the first, rather than a shared, and narrow, population bottleneck.

In our model we discriminate between changes in a population arising from selection and those arising due to the population bottleneck. This is achieved by considering regions of the genome between which recombination or reassortment has removed linkage disequilibrium between alleles (Figure 3.4; compare with Figure 3.1A). As transmission involves whole viruses, the bottleneck N^T is preserved between regions. Meanwhile, in the absence of epistasis, selection acting upon one region of the virus does not influence the composition of the population in other parts of the genome. As such, a calculation encompassing multiple parts of the genome can estimate both N^T and the influence of selection; in the figure the case of a loose population bottleneck, with selection acting upon the first region is preferred. A model selection process (Kass and Raftery 1995) is used to distinguish models of neutral transmission from evolution under selection (Figure 3.5).

3.2.3 Notation and Qualitative Overview

For clarity, I here define the notation utilised in the derivation of the basic model, which extends that set out in the previous chapter. The viral population is described as a set of haplotypes with associated frequencies that changes in time during a transmission event. Given a number of (possibly non-consecutive) loci of interest in the viral genome, the set of haplotypes $\boldsymbol{h} = \{h_i\}$ describes a set of sequences having specific nucleotides at these loci. Within a viral population of finite size, the number of viruses with each haplotype h_i is described by

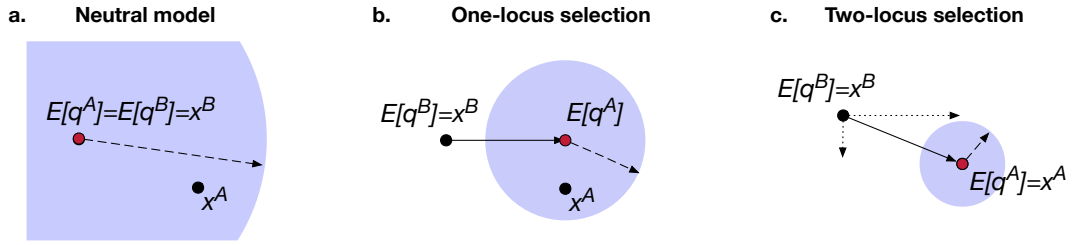


Figure 3.5. Models of neutrality and selection are compared, as illustrated in this simplified diagram. Black dots represent observations $\boldsymbol{x}^B$ and $\boldsymbol{x}^A$ while the red dot indicates the inferred expected position of $\boldsymbol{q}^A$. The solid line joining these (b,c) indicates the inferred action of selection, with dotted lines showing components of this vector (c). The blue circle represents the optimised variance in the position of $\boldsymbol{q}^A$; the length of its radius, shown as a dashed line, is inversely related to the bottleneck size. In the neutral case, the difference between observations is explained by the bottleneck alone. More complex models of selection fit $\boldsymbol{q}^A$ more closely to $\boldsymbol{x}^A$ and with reduced variance, giving higher inferred values of N^T.

the vector $\boldsymbol{n} = \{n_i\}$. Frequencies of each haplotype within the population are denoted by the vector $\boldsymbol{q} = \{q_i\}$, while observations of the population collected via sequencing are denoted by the vector $\boldsymbol{x} = \{x_i\}$, where x_i is the number of sampled viruses with haplotype h_i.

The transmission event is now described according to the framework outlined in Figure 3.2. A population of viruses $\boldsymbol{q}^B$ undergoes transmission with some bottleneck N^T, creating a founder population with haplotype frequencies $\boldsymbol{q}^F$ in the recipient. Selection influencing this transmission process is described by the function $S^T(\boldsymbol{q})$, which changes the frequency of haplotypes according to the relative propensity of each haplotype to transmit. For example, selection may favour the transmission of viruses containing a specific genetic variant, increasing the expected proportion of viruses with this variant in the founder population. Within the host, the viral population grows rapidly to create the population $\boldsymbol{q}^A$. During this growth process, genetic drift affects the population in a manner according to the effective population size N^G. Observations of the system are made via genome sequencing of samples collected before and after transmission, and are denoted $\boldsymbol{x}^B$ and $\boldsymbol{x}^A$ respectively; the total numbers of sequence reads in each are denoted N^B and N^A. Given the observations $\boldsymbol{x}^B$ and $\boldsymbol{x}^A$, we wish to estimate the size of the population bottleneck N^T and the extent of selection for transmissibility S^T.

Where I consider multiple replicate transmission events, I assume that each transmission has its own transmission bottleneck N^T; different numbers of viruses may infect different hosts. On the contrary, selection is assumed to operate

consistently between hosts; a variant which makes a virus grow more efficiently in one host does the same in another.

3.2.4 Likelihood Framework

By interpreting the transmission model in Figure 3.2 as a directed graphical model (Bishop 2006) we may consider the joint probability of the encircled variables given the external quantities. According to the principle of conditional probability we may write

$$\begin{aligned} &P(\boldsymbol{x}^B, \boldsymbol{x}^A, \boldsymbol{q}^B, \boldsymbol{q}^F, \boldsymbol{q}^A | C, N^T, N^G, S^T) \\ &\qquad = P(\boldsymbol{x}^B, \boldsymbol{x}^A, \boldsymbol{q}^F, \boldsymbol{q}^A | \boldsymbol{q}^B, C, N^T, N^G, S^T) P(\boldsymbol{q}^B) \end{aligned} \tag{3.1}$$

From the graphical model it is evident that $\boldsymbol{x}^B$ is conditionally independent of the post-transmission variables given $\boldsymbol{q}^B$. Mathematically, that is,

$$\left(\boldsymbol{x}^B \perp\!\!\!\perp \{\boldsymbol{x}^A, \boldsymbol{q}^F, \boldsymbol{q}^A\}\right) | \boldsymbol{q}^B \tag{3.2}$$

which implies that

$$\begin{aligned} &P(\boldsymbol{x}^B, \boldsymbol{x}^A, \boldsymbol{q}^B, \boldsymbol{q}^F, \boldsymbol{q}^A | C, N^T, N^G, S^T) \\ &\qquad = P(\boldsymbol{x}^B | \boldsymbol{q}^B, C) P(\boldsymbol{x}^A, \boldsymbol{q}^F, \boldsymbol{q}^A | \boldsymbol{q}^B, C, N^T, N^G, S^T) P(\boldsymbol{q}^B) \end{aligned} \tag{3.3}$$

The random variable $\boldsymbol{q}^B$ represents an unknown quantity that we necessarily need to determine in order to construct the likelihood framework. To this end we model $\boldsymbol{q}^B$ as a multivariate normal distributed random variable with mean $\boldsymbol{\mu}^B$ and covariance matrix Σ^B. This allows us to identify a multivariate normal probability density function with $P(\boldsymbol{q}^B)$, which for completeness we now refer to as $P(\boldsymbol{q}^B | \boldsymbol{\mu}^B, \Sigma^B)$. Noting that the true composition of the pre-transmission population is unknown, we average over all possible values of $\boldsymbol{q}^B$:

$$\begin{aligned} &P(\boldsymbol{x}^B, \boldsymbol{x}^A, \boldsymbol{q}^F, \boldsymbol{q}^A | C, N^T, N^G, S^T, \boldsymbol{\mu}^B, \Sigma^B) \\ &\quad = \int P(\boldsymbol{x}^B | \boldsymbol{q}^B, C) P(\boldsymbol{x}^A, \boldsymbol{q}^F, \boldsymbol{q}^A | \boldsymbol{q}^B, C, N^T, N^G, S^T) P(\boldsymbol{q}^B | \boldsymbol{\mu}^B, \Sigma^B) \mathrm{d}\boldsymbol{q}^B \end{aligned} \tag{3.4}$$

As all three probabilities depend on $\boldsymbol{q}^B$, they need to be jointly evaluated under the integral. In turn, this results in a joint likelihood for $\boldsymbol{\mu}^B$, Σ^B, N^T and S^T. Whilst technically possible to evaluate, this approach is computationally very intensive and results in long execution times. Instead, we use an approximation wherein $\boldsymbol{\mu}^B$ and Σ^B are initially inferred using only the before data and subsequently employed in the post-transmission framework with a view to inferring parameters of transmission (N^T and S^T). We may therefore define a likelihood for $\boldsymbol{\mu}^B$ and Σ^B

$$L(\boldsymbol{\mu}^B, \Sigma^B | \boldsymbol{x}^B, C) = \int P(\boldsymbol{x}^B | \boldsymbol{q}^B, C) P(\boldsymbol{q}^B | \boldsymbol{\mu}^B, \Sigma^B) \mathrm{d}\boldsymbol{q}^B \tag{3.5}$$

The first term in this likelihood, corresponding to the initial observation of the system, $\boldsymbol{x}^B$, represents a straightforward sampling of the system, drawing from a collection of viral haplotypes $\boldsymbol{q}^B$. Such a process can be modelled using a multinomial distribution. However, as is well known (Illingworth et al. 2017), next-generation sequence data are error-prone, such that less information is contained within the sample than would be contained in a multinomial sample of equivalent depth to the sample. A Dirichlet-multinomial distribution may be used to capture this reduction of information (Illingworth 2015, 2016), such that

$$P(\boldsymbol{x}^B | \boldsymbol{q}^B, C) = \frac{\Gamma(N^B + 1)}{\prod_i (x_i^B + 1)} \frac{\Gamma\left(\sum_i C q_i^B\right)}{\Gamma\left(\sum_i x_i^B + C q_i^B\right)} \prod_i \frac{\Gamma(x_i^B + C q_i^B)}{\Gamma(C q_i^B)} \tag{3.6}$$

where C, which alters the variance of the distribution, characterises the extent of noise in the data. The parameter C can be estimated given independent observations of identical parameters, such as haplotype or single allele frequencies; in the application to experimental data, time-resolved variant frequencies derived from the sequence data were used for this purpose (Illingworth 2015). As mentioned, $P(\boldsymbol{q}^B | \boldsymbol{\mu}^B, \Sigma^B)$ represents a multivariate normal in $\boldsymbol{\mu}^B$ and Σ^B. Given an estimate for C, we may therefore optimise Equation 3.5 for $\boldsymbol{\mu}^B$ and Σ^B.

Given maximum likelihood estimates for $\boldsymbol{\mu}^B$ and Σ^B, we now turn to the post-transmission part of the framework for which we have a joint probability of

$$P(\boldsymbol{x}^A, \boldsymbol{q}^F, \boldsymbol{q}^A | C, N^T, N^G, S^T, \boldsymbol{\mu}^B, \Sigma^B) = \int P(\boldsymbol{x}^A, \boldsymbol{q}^F, \boldsymbol{q}^A | \boldsymbol{q}^B, C, N^T, N^G, S^T) P(\boldsymbol{q}^B | \boldsymbol{\mu}^B, \Sigma^B) \mathrm{d}\boldsymbol{q}^B \tag{3.7}$$

Employing the rule of conditional probability repeatedly, we get

$$P(\boldsymbol{x}^A, \boldsymbol{q}^F, \boldsymbol{q}^A | C, N^T, N^G, S^T, \boldsymbol{\mu}^B, \Sigma^B) = P(\boldsymbol{x}^A | \boldsymbol{q}^A, C) P(\boldsymbol{q}^A | \boldsymbol{q}^F, N^G) \int P(\boldsymbol{q}^F | \boldsymbol{q}^B, N^T, S^T) P(\boldsymbol{q}^B | \boldsymbol{\mu}^B, \Sigma^B) \mathrm{d}\boldsymbol{q}^B \tag{3.8}$$

In the above, the expression $P(\boldsymbol{x}^A | \boldsymbol{q}^A, C)$ may be defined in a manner similar to Equation 3.6 dependent here upon the haplotype frequencies $\boldsymbol{q}^A$. The remaining parts of this equation can also be described as sampling events. A sample of the population in the donor animal transmits to the recipient, generating a founder population, $\boldsymbol{q}^F$. The founder population multiplies within the host, with offspring being sampled from the founder population to generate the final population $\boldsymbol{q}^A$. As $\boldsymbol{q}^F$ and $\boldsymbol{q}^A$ are both unknown variables, we need to marginalise them out:

$$P(\boldsymbol{x}^A | C, N^T, N^G, S^T, \boldsymbol{\mu}^B, \Sigma^B) = \sum_{\boldsymbol{q}^A} P(\boldsymbol{x}^A | \boldsymbol{q}^A, C) \sum_{\boldsymbol{q}^F} P(\boldsymbol{q}^A | \boldsymbol{q}^F, N^G) \int P(\boldsymbol{q}^F | \boldsymbol{q}^B, N^T, S^T) P(\boldsymbol{q}^B | \boldsymbol{\mu}^B, \Sigma^B) \mathrm{d}\boldsymbol{q}^B \tag{3.9}$$

The sums are over all the possible values that $\boldsymbol{q}^F$ and $\boldsymbol{q}^A$ may take. Despite representing frequencies, $\boldsymbol{q}^F$ and $\boldsymbol{q}^A$ do in fact take on discrete values as they arise on the basis of sampling events. Finally, if we consider N^G as a known quantity, something which will be justified shortly, as well as C, $\boldsymbol{\mu}^B$ and Σ^B, which are all independently inferred, we may define a likelihood for the transmission event:

$$L(N^T, S^T | \boldsymbol{x}^A, N^G, \boldsymbol{\mu}^B, \Sigma^B) = \sum_{\boldsymbol{q}^A} P(\boldsymbol{x}^A | \boldsymbol{q}^A, C) \sum_{\boldsymbol{q}^F} P(\boldsymbol{q}^A | \boldsymbol{q}^F, N^G) \int P(\boldsymbol{q}^F | \boldsymbol{q}^B, N^T, S^T) P(\boldsymbol{q}^B | \boldsymbol{\mu}^B, \Sigma^B) \mathrm{d}\boldsymbol{q}^B \tag{3.10}$$

I will go on to describe the calculation of both Equations 3.5 and 3.10, however, I will first define how selection is incorporated into our model.

3.2.4.1 Excursus: Modelling Selection

Within our model, the functions describing selection are potentially complex, each having a number of parameters equal to the number of haplotypes in the system. In common with previous approaches to studying within-host influenza evolution (Illingworth, Fischer and Mustonen 2014) we adopt a hierarchical model of selection whereby the fitness of a haplotype is calculated from a set of one- or multi-locus components, describing the advantage or disadvantage of a specific nucleotide, or nucleotides, at a single locus or set of loci. Model selection is then used to identify the most parsimonious explanation of the data.

Formally, we denote the j^{th} component of the haplotype h_i as h_{ij}, with $h_{ij} \in \{A, C, G, T\}$. In a fitness model, a parameter is defined as the pair of values (σ_k, g_k), where σ_k is a real number, denoting the difference in fitnesses of individuals with and without the allele (Kimura 1955), and g_k is a vector of components $g_{kj} \in \{A, C, G, T, -\}$ denoting the haplotypes to which this selection applies. We now define

$$g_k \cdot h_i = \prod_j g_{kj} \times h_{ij} \tag{3.11}$$

where

$$g_{kj} \times h_{ij} = \begin{cases} 1, & \text{if } g_{kj} = h_{ij} \\ 1, & \text{if } g_{kj} = - \\ 0, & \text{if } g_{kj} \neq -, g_{kj} \neq h_{ij} \end{cases} \tag{3.12}$$

The fitness of a haplotype h_i is then given as

$$w_i = \exp\left(\sum_k \sigma_k (g_k \cdot h_i)\right) \tag{3.13}$$

where the sum is calculated over all fitness parameters k. To give an example, a single-locus fitness parameter would have a single element of g_k that was either A, C, G, or T. Supposing this element to be at position j, it would convey the fitness advantage σ_k to all haplotypes with the given nucleotide at position j in the genome.

3.2.4.2 Selection in a Transmission Event

Selection is incorporated into the transmission event from donor to recipient by representing this event as a biased sampling process. As we are not considering data here, noise is not an issue. I therefore model the population $\boldsymbol{q}^F$ as arising via a multinomial sampling process of depth N^T from a set of genotypes with frequencies $S^T(\boldsymbol{q}^B)$, where S^T represents the role of selection in the transmission event. We write

$$P(\boldsymbol{q}^F|\boldsymbol{q}^B, N^T, S^T) = \frac{N^T!}{\prod_i n_i^F!}\prod_i (S^T(\boldsymbol{q}^B))_i^{n_i^F} \tag{3.14}$$

where

$$\left(S^T(\boldsymbol{q}^B)\right)_i = \frac{w_i^T q_i^B}{\sum_{i'} w_{i'}^T q_{i'}^B} \tag{3.15}$$

defines a distorted population based on the haplotype fitnesses $\boldsymbol{w}^T = \{w_i^T\}$, representing the relative propensity of each haplotype h_i for transmission. We note here that $q_i^F = \frac{n_i^F}{N^T}$, where the vector $\boldsymbol{n}^F$ describes the number of copies of each haplotype in the founder population. This construction makes the assumption that selection acts immediately prior to the bottleneck event itself.

3.2.4.3 Within-Host Growth

Concerning genetic drift, we note that the number of viruses in a host grows rapidly, with experiments suggesting that a single infected cell can produce between 10^3 and 10^4 viruses (Stray and Air 2001). However not every such virus is viable, and one estimate has put the number of naive cells infected by an infected influenza cell at 22 (Baccam et al. 2006). I here approximate the within-host growth of the virus as a single multinomial draw, compressing growth to a single round of sampling, with the variance effective population size $N^G = gN^T$. By default I set the growth factor g to be equal to 22. This approach is distinct from the branching process used in another estimate of bottleneck size (Sobel Leonard et al. 2017a); our assumption that viruses infect different cells in the host, with competition between viruses occurring after the release of viruses from cells, leads to a Wright-Fisher-type population model, in which the rapid growth of the viral population leads to a smaller amount of genetic drift than inferred in that model.

3.2.4.4 Approximation of the Likelihood Function

I now turn to calculating the likelihood functions of Equations 3.5 and 3.10. In the case of Equation 3.10, the likelihood for the transmission parameters, computing this expression turns out to be effectively impossible under certain circumstances. This is due to the summation over $\boldsymbol{q}^F$ and $\boldsymbol{q}^A$, where the number of possible outcomes grows combinatorially with N^T and the number of haplotypes, in turn making this calculation intractable. Instead I consider a continuous approximation in which the random variables of the model (Figure 3.2) are represented by multivariate normal distributions, each defined by a mean and covariance matrix. As a result, the summations over $\boldsymbol{q}^F$ and $\boldsymbol{q}^A$ in Equation 3.10 are replaced with integrals:

$$\begin{aligned} L(N^T, S^T|\boldsymbol{x}^A) = \int P(\boldsymbol{x}^A|\boldsymbol{q}^A, C) \Bigg[\int P(\boldsymbol{q}^A|\boldsymbol{q}^F, N^G) \\ \left[\int P(\boldsymbol{q}^F|\boldsymbol{q}^B, N^T, S^T) P(\boldsymbol{q}^B|\boldsymbol{\mu}^B, \Sigma^B) \mathrm{d}\boldsymbol{q}^B \right] \mathrm{d}\boldsymbol{q}^F \Bigg] \mathrm{d}\boldsymbol{q}^A \end{aligned} \tag{3.16}$$

By ignoring higher order moments, we may then calculate the individual components of the system by appealing to a moments based approach for the evaluation of integrals arising from marginalisation over unknown variables. This step follows multiple previous approaches to time-resolved data, in which moments-based approximations have been used to simplify the propagation of evolutionary models (Lacerda and Seoighe 2014; Tataru et al. 2017; Terhorst, Schlötterer and Song 2015; Tran, Hofrichter and Jost 2014).

I now outline the conditional moments required for calculating the likelihoods via the moments based approach. Firstly, given a sampling depth N^B and a dispersion parameter C, we describe $\boldsymbol{x}^B$ as a distribution with mean and variance derived from the Dirichlet-multinomial (Mosimann 1962):

$$\mathrm{E}\left[\boldsymbol{x}^B|\boldsymbol{q}^B\right] = N^B \boldsymbol{q}^B \tag{3.17}$$

and

$$\mathrm{var}\left[\boldsymbol{x}^B|\boldsymbol{q}^B\right] = \left(\frac{N^B + C}{1 + C}\right) N^B \left(\mathrm{Diag}(\boldsymbol{q}^B) - \boldsymbol{q}^B(\boldsymbol{q}^B)^\dagger\right) \equiv \beta N^B M(\boldsymbol{q}^B) \tag{3.18}$$

where $\beta = \left(\frac{N^B+C}{1+C}\right)$, $M(\boldsymbol{q}) = \text{Diag}(\boldsymbol{q}) - \boldsymbol{q}\boldsymbol{q}^\dagger$ and $\dagger$ indicates the transpose function.

The founder population $\boldsymbol{q}^F$ is sampled from $\boldsymbol{q}^B$. Its mean is given by the expression

$$\text{E}[\boldsymbol{q}^F|\boldsymbol{q}^B] = S^T(\boldsymbol{q}^B) \tag{3.19}$$

and its variance by

$$\text{var}[\boldsymbol{q}^F|\boldsymbol{q}^B] = \frac{1}{N^T}\left(\text{Diag}(S^T(\boldsymbol{q}^B)) - S^T(\boldsymbol{q}^B)S^T(\boldsymbol{q}^B)^\dagger\right) \equiv \frac{1}{N^T}M(S^T(\boldsymbol{q}^B)) \tag{3.20}$$

arising from a multinomial sample of depth N^T and the selectively shifted frequencies $S^T\left(\boldsymbol{q}^B\right)$.

Similarly, the within-host growth process may be represented by a distribution with mean $\text{E}[\boldsymbol{q}^A|\boldsymbol{q}^F] = \boldsymbol{q}^F$ and variance $\text{var}[\boldsymbol{q}^A|\boldsymbol{q}^F] = \frac{1}{N^G}M(\boldsymbol{q}^F)$. As for the pre-transmission case, a Dirichlet-multinomial likelihood with sampling depth N^A, haplotype frequencies $\boldsymbol{q}^A$ and dispersion parameter C may be used to model the sequencing of the population post-transmission. This distribution can be approximated as a multivariate normal with mean

$$\text{E}[\boldsymbol{x}^A|\boldsymbol{q}^A] = N^A\boldsymbol{q}^A \tag{3.21}$$

and variance

$$\text{var}[\boldsymbol{x}^A|\boldsymbol{q}^A] = \left(\frac{N^A+C}{1+C}\right)N^AM(\boldsymbol{q}^A) \equiv \alpha N^AM(\boldsymbol{q}^A) \tag{3.22}$$

where $\alpha = \left(\frac{N^A+C}{1+C}\right)$ is defined for notational convenience.

Having established the above distributions, we are now equipped to carry out the relevant marginalisations (Equations 3.5 and 3.16) using the law of total expectation and the law of total variance. Starting with the pre-transmission compound distribution, the marginalisation over $\boldsymbol{q}^B$ yields a mean of

$$\text{E}[\boldsymbol{x}^B] = \text{E}[\text{E}[\boldsymbol{x}^B|\boldsymbol{q}^B]] = \text{E}[N^B\boldsymbol{q}^B] = N^B\boldsymbol{\mu}^B \tag{3.23}$$

and a variance of

$$\begin{aligned}\mathrm{var}(\boldsymbol{x}^B) &= \mathrm{E}[\mathrm{var}[\boldsymbol{x}^B|\boldsymbol{q}^B]] + \mathrm{var}[\mathrm{E}[\boldsymbol{x}^B|\boldsymbol{q}^B]] \\ &= \mathrm{E}\left[\beta N^B \left(\mathrm{Diag}(\boldsymbol{q}^B) - \boldsymbol{q}^B(\boldsymbol{q}^B)^\dagger\right)\right] + \mathrm{var}[N^B\boldsymbol{q}^B] \\ &= \beta N^B \left(\mathrm{Diag}(\mathrm{E}[\boldsymbol{q}^B]) - \mathrm{E}[\boldsymbol{q}^B]E[\boldsymbol{q}^B]^\dagger\right) + N^B\left(N^B - \beta\right)\mathrm{var}[\boldsymbol{q}^B] \\ &= \beta N^B M\left(\boldsymbol{\mu}^B\right) + N^B\left(N^B - \beta\right)\Sigma^B\end{aligned} \tag{3.24}$$

These expressions characterise the pre-transmission compound distribution from Equation 3.5 in terms of a normal distribution. We identify values of $\boldsymbol{\mu}^B$ and Σ^B maximising this likelihood. The matrix Σ^B has dimensionality k^2 where k is the number of haplotypes in the system, a number which may potentially be large. Accurately determining so many parameters from the available data is unrealistic. In preference to obtaining an ill-defined covariance matrix we make the approximation that the off-diagonal elements of Σ^B are zero, i.e. we disregard between-haplotype correlations in specifying the uncertainty in $\boldsymbol{\mu}^B$. I note that ignoring the variance component altogether results in an underestimation of the population bottleneck (see Section 3.3.2 and Figure 3.9).

Moving on to the post-transmission process, the marginalisation over $\boldsymbol{q}^B$ results in a mean of

$$\mathrm{E}[\boldsymbol{q}^F] = \mathrm{E}[\mathrm{E}[\boldsymbol{q}^F|\boldsymbol{q}^B]] = \mathrm{E}[S^T(\boldsymbol{q}^B)] \approx S^T(\mathrm{E}[\boldsymbol{q}^B]) = S^T(\boldsymbol{\mu}^B) \tag{3.25}$$

where in the penultimate step we used the first-order second-moment approximation to a vector function acting on a random variable[1]. The law of total variance yields

$$\begin{aligned}\mathrm{var}(\boldsymbol{q}^F) &= \mathrm{E}[\mathrm{var}[\boldsymbol{q}^F|\boldsymbol{q}^B]] + \mathrm{var}[\mathrm{E}[\boldsymbol{q}^F|\boldsymbol{q}^B]] \\ &= \mathrm{E}\left[\frac{1}{N^T}M(S^T(\boldsymbol{q}^B))\right] + \mathrm{var}\left[S^T(\boldsymbol{q}^B)\right] \\ &= \frac{1}{N^T}M(\mathrm{E}[S^T(\boldsymbol{q}^B)]) + \left(1 - \frac{1}{N^T}\right)\mathrm{var}[S^T(\boldsymbol{q}^B)] \\ &\approx \frac{1}{N^T}M(S^T(\mathrm{E}[\boldsymbol{q}^B])) + \left(1 - \frac{1}{N^T}\right)\left(DS^T\big|_{\mathrm{E}[\boldsymbol{q}^B]}\right)\mathrm{var}[\boldsymbol{q}^B]\left(DS^T\big|_{\mathrm{E}[\boldsymbol{q}^B]}\right)^\dagger \\ &= \frac{1}{N^T}M(S^T(\boldsymbol{\mu}^B)) + \left(1 - \frac{1}{N^T}\right)\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)\Sigma^B\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger\end{aligned} \tag{3.26}$$

[1]The extension from scalar-valued functions to vector-valued functions for the first-order second-moment method is straightforward, see Appendix F.

Note that $(DS)_i^j = \frac{\partial S_i}{\partial q_j}$ is the Jacobian matrix arising from the first-order second-moment approximation.

Marginalisation over $\boldsymbol{q}^F$ yields a mean of

$$\mathrm{E}[\boldsymbol{q}^A] = \mathrm{E}[\mathrm{E}[\boldsymbol{q}^A|\boldsymbol{q}^F]] = \mathrm{E}[\boldsymbol{q}^F] = S^T(\boldsymbol{\mu}^B) \tag{3.27}$$

and variance

$$\begin{aligned}
\mathrm{var}(\boldsymbol{q}^A) &= \mathrm{E}[\mathrm{var}[\boldsymbol{q}^A|\boldsymbol{q}^F]] + \mathrm{var}[\mathrm{E}[\boldsymbol{q}^A|\boldsymbol{q}^F]] \\
&= \mathrm{E}\left[\frac{1}{N^G}\left(\mathrm{Diag}(\boldsymbol{q}^F) - \boldsymbol{q}^F(\boldsymbol{q}^F)^\dagger\right)\right] + \mathrm{var}[\boldsymbol{q}^F] \\
&= \frac{1}{N^G}\left(\mathrm{Diag}(\mathrm{E}[\boldsymbol{q}^F]) - \mathrm{E}[\boldsymbol{q}^F]E[\boldsymbol{q}^F]^\dagger\right) + \left(1 - \frac{1}{N^G}\right)\mathrm{var}[\boldsymbol{q}^F] \\
&= \frac{1}{N^G}M\left(S^T(\boldsymbol{\mu}^B)\right) + \left(1 - \frac{1}{N^G}\right)\left(\frac{1}{N^T}M(S^T(\boldsymbol{\mu}^B)) + \right. \\
&\qquad \left. \left(1 - \frac{1}{N^T}\right)\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)\Sigma^B\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger\right) \\
&= \frac{N^T + N^G - 1}{N^T N^G}M\left(S^T(\boldsymbol{\mu}^B)\right) + \\
&\qquad \frac{N^T N^G - N^T - N^T + 1}{N^T N^G}\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)\Sigma^B\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger \\
&\equiv \gamma M\left(S^T(\boldsymbol{\mu}^B)\right) + \delta\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)\Sigma^B\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger
\end{aligned} \tag{3.28}$$

where in the last step we defined $\gamma = \left(\frac{N^T+N^G-1}{N^T N^G}\right)$ and $\delta = \frac{N^T N^G - N^T - N^G + 1}{N^T N^G}$.

Treating the integral over $\boldsymbol{q}^A$ in a similar manner, we obtain by the law of total expectation

$$\mathrm{E}[\boldsymbol{x}^A] = \mathrm{E}[\mathrm{E}[\boldsymbol{x}^A|\boldsymbol{q}^A]] = \mathrm{E}[N^A\boldsymbol{q}^A] = N^A\mathrm{E}[\boldsymbol{q}^A] = N^A S^T\left(\boldsymbol{\mu}^B\right) \tag{3.29}$$

Analogously, the law of total variance yields

$$
\begin{aligned}
\mathrm{var}(\boldsymbol{x}^A) &= \mathrm{E}[\mathrm{var}[\boldsymbol{x}^A|\boldsymbol{q}^A]] + \mathrm{var}[\mathrm{E}[\boldsymbol{x}^A|\boldsymbol{q}^A]] \\
&= \mathrm{E}\left[\alpha N^A M\left(\boldsymbol{q}^A\right)\right] + \mathrm{var}[N^A\boldsymbol{q}^A] \\
&= \alpha N^A\left(\mathrm{Diag}(\mathrm{E}\left[\boldsymbol{q}^A\right] - \mathrm{E}\left[\boldsymbol{q}^A\right]\mathrm{E}\left[\boldsymbol{q}^A\right]^\dagger\right) + N^A\left(N^A-\alpha\right)\mathrm{var}[\boldsymbol{q}^A] \\
&= \alpha N^A M(S^T(\boldsymbol{\mu}^B)) \\
&\qquad + N^A\left(N^A-\alpha\right)\left(\gamma M\left(S^T(\boldsymbol{\mu}^B)\right) + \delta\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)\Sigma^B\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger\right) \\
&= N^A\left(\alpha + (N^A-\alpha)\gamma\right)M(S^T(\boldsymbol{\mu}^B)) \\
&\qquad + N^A\left(N^A-\alpha\right)\delta\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)\Sigma^B\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger
\end{aligned}
\tag{3.30}
$$

The above expressions represent mean and covariance matrices of a multivariate normal distribution resulting from the evaluation of marginalisations in Equation 3.16. As such, this defines a likelihood for the transmission event, from which, given the data $\boldsymbol{x}^A$ and our estimates for μ^B and Σ^B, we may infer maximum likelihood values for N^T and S^T.

3.2.4.5 Excursus: A Note on Dimensionality

In the above I considered multivariate normal distributions with means and variances defined by multinomial sampling processes. In general a multinomial distribution with frequencies $\boldsymbol{q}$ and sampling depth N may be approximated by a normal distribution as follows:

$$\mathrm{Multi}(N, \boldsymbol{q}) \approx \mathcal{N}(N\boldsymbol{q}, NM) \tag{3.31}$$

where $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ denotes the multivariate normal distribution, $M = \mathrm{Diag}(\boldsymbol{q}) - \boldsymbol{q}\boldsymbol{q}^\dagger$ is a square matrix and $\dagger$ indicates vector transpose. This approximation holds in the regime where N is large and $\boldsymbol{q}$ sufficiently far from the boundaries (Geyer 2006).

By construction, a multinomial sample $\boldsymbol{x}$ is constrained by the requirement that $\sum_{i=1}^k x_i = N$. Effectively, this implies that a multinomial random variable of dimension k is actually of dimension $k-1$. This has implications for the normal approximation (Equation 3.31), as the probability density function for the multivariate normal distribution is only non-degenerate when the covariance matrix Σ is positive-definite. To ensure this I reduce the dimensionality by one

when evaluating likelihoods based on multivariate normal distributions. In the remainder of this work, the notation will be overloaded such that e.g. $\boldsymbol{x}$ and $\boldsymbol{q}$ denote both reduced and complete sets.

The dimensional reduction is furthermore a requirement for ensuring the correct behaviour of the selection function in the limit of no selection. In particularly, we expect the variance in Equation 3.30 to collapse to the variance of a neutral model (see Appendix C) as $\left(S^T(\boldsymbol{q}^B)\right)_i \to \boldsymbol{q}_i^B$ (or equivalently, $w_i \to 1$) for all i:

$$\mathrm{var}(\boldsymbol{x}^A) = N^A \left(\alpha + \left(N^A - \alpha\right)\gamma\right) M\left(\boldsymbol{\mu}^B\right) + N^A\left(N^A - \alpha\right)\delta\Sigma^B \tag{3.32}$$

Clearly this is only the case if the Jacobian matrix correspondingly approaches the identity matrix, i.e. $DS^T\big|_{\boldsymbol{\mu}^B} \to \mathbb{1}$. In the full-dimensional framework, the Jacobian matrix is defined as:

$$(DS)_i^j = \frac{\partial S(\boldsymbol{q})_i}{\partial q_j} = \frac{\partial}{\partial q_j}\frac{w_i q_i}{\sum_{i'} w_{i'} q_{i'}} = \begin{cases} \frac{w_i}{\sum_{i'} w_{i'} q_{i'}} - \frac{w_i^2 q_i}{\left(\sum_{i'} w_{i'} q_{i'}\right)^2} & \text{if } i = j \\ -\frac{w_i w_j q_i}{\left(\sum_{i'} w_{i'} q_{i'}\right)^2} & \text{if } i \neq j \end{cases} \tag{3.33}$$

which does not collapse to the identity matrix as $w_i \to 1$. On the other hand, in the reduced dimensionality framework the selection function takes form

$$S(\boldsymbol{q})_i = \frac{w_i q_i}{w_1 q_1 + w_2 q_2 + \ldots + w_{K-1} q_{K-1} + w_K(1 - q1 - q2 - \ldots - q_{K-1})} \equiv \frac{w_i q_i}{\eta} \tag{3.34}$$

for K haplotypes where we defined the denominator as η. Without loss of generality we have chosen to express the Kth haplotype in terms of the preceding $K-1$ haplotypes. Under this setup, the Jacobian matrix takes the form

$$(DS)_i^j = \begin{cases} \frac{w_i}{\eta} - \frac{w_i q_i}{\eta^2}\left(w_i - w_K\right) & \text{if } i = j \\ -\frac{w_i q_i}{\eta^2}\left(w_j - w_K\right) & \text{if } i \neq j \end{cases} \tag{3.35}$$

which appropriately collapses to the identity matrix as $w_i \to 1$:

$$\lim_{w_i \to 1} (DS)_i^j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{3.36}$$

3.2.5 Maximum Likelihood Optimisation Method for Transmission

In order to identify maximum likelihood parameters of transmission, a hill climbing optimisation method was employed. This optimisation approach was dynamic in nature, utilising two variables, δ and n_{update}, to determine the scope and duration of the optimisation routine. The first parameter, δ, was initialised to unity and responsible for tracking convergence, this being defined as $\delta < 0.01$. Furthermore, δ represented the magnitude of changes occurring during an update step, details of which will be given in Sections 3.2.5.2 and 3.2.5.3 below. The second parameter, n_{update}, was initialised to a value of 100 update rounds and specified the number of update iterations prior to a recalculation of δ and n_{update} itself. Recalculation was based upon an acceptance rate r defined by the fraction of successful updates to the total number of attempted updates. Specifically, δ was updated to $\delta_{\text{new}} = \delta(0.90 + r)$, i.e. δ increased in size if the acceptance rate was larger than 10% and decreased otherwise. The underlying rationale being that a high acceptance rate represents a low degree of convergence which in turn warrants a higher value of δ in order to explore more distant regions of optimisation space. To control the optimisation process, an upper limit of $\delta = 5$ was enforced. Similarly, n_{update} was increased by a value of 10 if $r > 0.1$ and otherwise decreased by an identical amount. n_{update} was constrained on $[10, 500)$. Taken together, this defined a dynamic optimisation process ensuring adequate sampling of optimisation regions based on acceptance rates.

3.2.5.1 General Update Dynamics

General update dynamics were based around a set of acceptance rates. In addition to the overall acceptance rate, r, additional acceptance rates for bottleneck updates, r_{NT}, and for updates to selection coefficients, r_{sel}, were employed. The ratio of these, $f = \frac{r_{NT}}{r_{\text{sel}}}$, was used to define the probability of performing a bottleneck update; a uniform random number q, defined on [0,1), was generated and a bottleneck update performed if $q < f$, otherwise a selection update was carried out. This setup allowed for a disproportionate amount of updates of one kind or the other, which ensured a more rapid convergence. As mentioned above, acceptance rates and f were recalculated every n_{update} rounds.

3.2.5.2 Updating Bottleneck Values

Upon initiation, the transmission bottleneck size was set to $N^T = 100$. The magnitude of bottleneck updates was defined by δ and a factor c, which took values of 1, 5, 10 and 50 depending on whether the bottleneck was below 100, 500, 1000 and 10,000 respectively. Specifically, the updated bottleneck took a value of $N^T_{\text{new}} = N^T \pm \text{ceil}(\delta)c$ where ceil(x) returns the least integer greater than or equal to x and the direction of the update chosen at random.

3.2.5.3 Updating Selection Coefficients

At the beginning of the optimisation process, selection coefficients were initialised to zero. During an update, a selection coefficient was chosen at random and it's magnitude changed to $s_{\text{new}} = s + (q - 0.5) * \delta$ where, as before, q is a uniform random number on [0,1). Selection coefficients were limited to $s = \pm 10$.

3.2.6 Maximum Likelihood Optimisation Method for q^B

For the inference of maximum likelihood values of μ^B and Σ^B a highly similar approach was taken, albeit with a few subtle differences. Here, δ was initialised to unity and convergence defined as $\delta < 10^{-6}$ whilst n_{update} was initialised to 1000 update rounds. Given acceptance rate r, the parameters were updated to $\delta_{\text{new}} = \delta(0.95 + r)$ and $n_{\text{update,new}} = \text{ceil}(n_{\text{update}}(0.95 + r))$ every n_{update} rounds. An upper limit of $\delta < 10$ was enforced whilst n_{update} was restricted to the region $[100, 1000]$. Collectively, this heuristically derived setup guaranteed a rapid inference of the components of the pre-transmission population.

3.2.6.1 Update Dynamics

A simpler approach was taken to the inference of μ^B and Σ^B than was the case for transmission. Initially, a k-dimensional μ^B was initialised to $\mu^B_i = 1/k$ whilst Σ^B was set to $\Sigma^B = \text{Diag}(a)$. We here employed a relatively large value of $a = 0.1$, which generally represented a highly unlikely starting point, as this guaranteed a defined move away from the initial state during optimisation. Conversely, using a smaller value, such as $a = 10^{-10}$, often resulted in an optimisation process getting stuck in a local maximum.

During the update routine, either μ^B or Σ^B was chosen for update, this decision being made at random. For updating μ^B, a random integer i on $[0, k]$ was generated, designating the haplotype being altered. Next, a uniform random

number q on $[0, 1)$ was used for updating the ith entry of μ^B, namely $\mu^B_{i,\text{new}} = \mu^B_i + \delta(q - 0.5)$ with the constraint that $\mu^B_{i,\text{new}} > 10^{-11}$. Finally, μ^B was scaled to have unit sum. Similarly, for updating Σ^B the (i, i) entry assumed a value of $\Sigma^B_{i,i,\text{new}} = \Sigma^B_{i,i} + \delta(q - 0.5)$ where $\Sigma^B_{i,i,\text{new}} > 10^{-11}$.

3.2.7 Reversion to a Discrete Likelihood Function

Given a mean and covariance matrix for the likelihood function, we can approximate the likelihood by the probability density function of a multivariate normal distribution. However, where the variance of this distribution is very small in one dimension, as can occur under an inference of very strong selection, the density function evaluated at a point can become arbitrarily large. For this reason a Gaussian cubature approach was used to calculate the integral of the final likelihood over the unit cube described by each observation $\boldsymbol{x}$, once optimisation had been completed. Approximate numerical integrals were calculated using the software package cubature (S. G. Johnson n.d.).

We also note that under neutrality it is possible to derive a discrete compound solution, i.e. a solution in which the likelihood is evaluated using a Dirichlet-multinomial rather than a multivariate normal distribution. Such a solution represents an improvement on the normal approximation by taking into account skewness. The Dirichlet-multinomial is defined by a probability vector $\boldsymbol{p}$ and a parameter vector $\boldsymbol{\alpha}$. In Appendix A we present a derivation for obtaining $\boldsymbol{\alpha}$ based on the compound mean $\boldsymbol{\mu}$ and the covariance matrix Σ. In Appendix B we have briefly examined the performance of the discrete compound solution in a one-dimensional setting, however, it won't be explored further in this work.

3.2.8 Extension to Partial Haplotype Data

In the calculations above I made the implicit assumption that the observations $\boldsymbol{x}^B$ and $\boldsymbol{x}^A$ consist of sets of complete viral haplotypes $\boldsymbol{h}_i$. However, short-read sequencing technologies generally result in sets of individual reads which only cover a subset of the genetic loci of interest; I here refer to these reads as partial haplotypes. In this framework the data represent direct observations of partial haplotypes in the set $\boldsymbol{h}^P = \{\boldsymbol{h}^{\text{P}}_1, \ldots, \boldsymbol{h}^{\text{P}}_L\}$, where each of the sets $\boldsymbol{h}^{\text{P}}_l$ is a vector of haplotypes spanning a common subset of the loci spanned by the full haplotypes in $\boldsymbol{h}$. Population-wide observations of these partial haplotypes are then defined by $\boldsymbol{x}^{\text{P}} = \{\boldsymbol{x}^{\text{P}}_1, \ldots, \boldsymbol{x}^{\text{P}}_L\}$ with $\boldsymbol{x}^{\text{P}}_l = \{x^{\text{P}}_{li}\}$ where x^{P}_{li} is the number of

reads with haplotype $\boldsymbol{h}_{li}^{\mathrm{P}}$. As a result, the total number of observations must now be defined on the basis of each set of partial haplotypes, e.g. $N_l^{B,\mathrm{P}} = \sum_i x_{li}^{\mathrm{P}}$ is the total number of observations of partial haplotypes in the set l. As each set of partial haplotype observations is independent of the others, we may reconstruct Equation 3.16 in the following terms:

$$\log L(N^T, S^T | \boldsymbol{x}^A, N^G, \boldsymbol{\mu}^B, \Sigma^B) = \sum_l \log L(N^T, S^T | \boldsymbol{x}_l^{A,\mathrm{P}}, N^G, \boldsymbol{\mu}^B, \Sigma^B) \quad (3.37)$$

Within this construction, bottleneck sizes and selection are conserved between partial haplotype sets, being evaluated at the full haplotype level. Each set of partial haplotype observations $\boldsymbol{x}_l^{\mathrm{P}}$ is considered as a sample drawn from a set of partial haplotypes with frequencies $\boldsymbol{q}_l^{\mathrm{P}}$, these frequencies being defined via a linear transformation of the full haplotype frequencies with matrix T_l. For example, given the full haplotypes $\{\mathrm{AG}, \mathrm{AT}, \mathrm{CG}, \mathrm{CT}\}$ and a set of partial haplotypes $\{\mathrm{A\text{-}}, \mathrm{C\text{-}}\}$, we have

$$\boldsymbol{q}_l^{\mathrm{P}} = T_l \boldsymbol{q} \quad (3.38)$$

or more explicitly,

$$\begin{pmatrix} q_{l1}^{\mathrm{P}} \\ q_{l2}^{\mathrm{P}} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{pmatrix} \quad (3.39)$$

Thus, as described above, the calculation of transmission and within-host growth under selection can be performed at the level of full haplotypes, switching into partial haplotype space only to evaluate the likelihoods of the observations. Re-deriving the results of Equations 3.23 and 3.24 for short-read sequence data, we find that the compound distribution for the $\boldsymbol{x}^B$ component has mean

$$\mathrm{E}[\boldsymbol{x}_l^{B,\mathrm{P}}] = N_l^{B,\mathrm{P}} T_l \, \boldsymbol{\mu}^B \quad (3.40)$$

and variance

$$\mathrm{var}(\boldsymbol{x}_l^{B,\mathrm{P}}) = \beta_l N_l^{B,\mathrm{P}} M\left(T_l \boldsymbol{\mu}^B\right) + N_l^{B,\mathrm{P}}\left(N_l^{B,\mathrm{P}} - \beta_l\right) T_l \Sigma^B T_l^\dagger \tag{3.41}$$

Similarly, for the $\boldsymbol{x}^A$ component of the likelihood, we get a mean of

$$\mathrm{E}[\boldsymbol{x}_l^{A,\mathrm{P}}] = N_l^{A,\mathrm{P}} T_l S^T\left(\boldsymbol{\mu}^B\right) \tag{3.42}$$

and variance

$$\begin{aligned}\mathrm{var}(\boldsymbol{x}_l^{A,\mathrm{P}}) = {} & N_l^{A,\mathrm{P}}\left(\alpha_l + (N_l^{A,\mathrm{P}} - \alpha_l)\gamma\right) M(T_l S^T(\boldsymbol{\mu}^B)) \\ & + N_l^{A,\mathrm{P}}\left(N_l^{A,\mathrm{P}} - \alpha_l\right)\delta T_l \left(DS^T\big|_{\boldsymbol{\mu}^B}\right)\Sigma^B\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger T_l^\dagger\end{aligned} \tag{3.43}$$

In Equation 3.43 we used the identity $T_l \mathrm{Diag}(\boldsymbol{\mu}^B) T_l^\dagger = \mathrm{Diag}(T_l \boldsymbol{\mu}^B)$ which is true if and only if T_l consists of zeroes and ones and if every column of T_l contains a single non-zero element, i.e. if the partial haplotype sets are independent from one another. A derivation of this identity is given in Appendix E and an application of it can be found in Appendix C.

3.2.9 Data From Multiple Genes

The mathematical framework outlined above utilises the haplotype information inherent to the data, and accounts for the effect of noise in the sequencing process (Figure 3.1B,C). However, in order to discriminate between changes in viral diversity arising from bottlenecking and selection (Figure 3.1A) it is necessary to consider data from different regions of the genome at which genetic diversity is nominally statistically independent. At high doses of influenza virus reassortment occurs rapidly, as has been observed both *in vitro* and in small animal infections (Marshall et al. 2013; Tao, Steel and Lowen 2014). In our analysis, distinct viral segments were therefore considered to be independent of one another in this manner, albeit sharing a common transmission bottleneck N^T, each transmitted virus being assumed to contain one of each viral segment. As such the likelihood in Equation 3.37 becomes

$$\log L(N^T, S^T|\boldsymbol{x}^A, N^G, \boldsymbol{\mu}^B, \Sigma^B) = \sum_m \sum_l \log L(N^T, S_m^T|\boldsymbol{x}_{ml}^{A,\mathrm{P}}, N^G, \boldsymbol{\mu}_m^B, \Sigma_m^B) \tag{3.44}$$

where the subscript m denotes information particular to a specific genomic region.

3.2.10 Data From Multiple Replicates

Replicate data are highly valuable for evolutionary inference (Achaz et al. 2014; Kofler and Schlötterer 2014). Within our calculation they provide an additional level of abstraction to the inference process. Under this framework I assumed that replicates share a common fitness landscape, S^T, whilst exhibiting individual bottleneck values. The validity of imposing a joint fitness landscape has been previously verified using data from a human challenge study (Sobel Leonard et al. 2017a). As a result, the likelihood from Equation 3.44 becomes

$$\log L(N^T, S^T|\boldsymbol{x}^A, N^G, \boldsymbol{\mu}^B, \Sigma^B) = \sum_r \sum_m \sum_l \log L(N_r^T, S_m^T|\boldsymbol{x}_{rml}^{A,\mathrm{P}}, N_r^G, \boldsymbol{\mu}_{rm}^B, \Sigma_{rm}^B) \tag{3.45}$$

where the subscript r denotes information particular to a specific replicate.

3.2.11 Implementation of Sobel Leonard et al. Method

For comparison of bottleneck estimates with existing methods I implemented the exact version of the beta-binomial inference scheme of Sobel Leonard et al. (Sobel Leonard et al. 2017b). The likelihood function for site i was defined as

$$L(N^T)_i = \sum_{j=0}^{N^T} P_{\text{beta-bin}}\left(x_{i,\text{minor}}^{A,\mathrm{SL}}|N_i^{A,\mathrm{SL}}, j, N^T - j\right) P_{\text{bin}}\left(j|N^T, q_{i,\text{minor}}^{B,\mathrm{SL}}\right) \tag{3.46}$$

where N^T is the bottleneck size, $P_{\text{beta-bin}}$ is the beta-binomial probability mass function, $x_{i,\text{minor}}^{A,\mathrm{SL}}$ is the number of recipient observations for the minor allele at site i, $N_i^{A,\mathrm{SL}}$ is the total number of recipient observations for site i, i.e. $N_i^{A,\mathrm{SL}} =$

$x^{A,\mathrm{SL}}_{i,\mathrm{minor}} + x^{A,\mathrm{SL}}_{i,\mathrm{major}}$, P_{bin} is the binomial probability mass function, and $q^{B,\mathrm{SL}}_{i,\mathrm{minor}}$ is the donor frequency for the minor allele at site i. Noting that the beta-binomial is undefined for $j = 0$ and $j = N^T$, I defined $j = 10^{-10}$ and $j = N^T - 10^{-10}$ respectively in these cases. The original authors did not discuss this further. I did not make use of the cumulative version of the likelihood function (Sobel Leonard et al. 2017b) as I avoid the problem of variant calling by fixing the number of required polymorphic loci when simulating data. The total likelihood for each bottleneck value was computed as

$$L(N^T) = \sum_{i=0}^{n_{\mathrm{sites}}} L(N^T)_i \tag{3.47}$$

where n_{sites} is the number of variant loci. Bottleneck inference was defined as the bottleneck associated with the largest likelihood value.

3.2.12 Generation of Simulated Data

Simulated data were generated in order to nominally reflect data from an influenza transmission event. As such, a single transmission event was modelled as the transmission of viruses each with eight independent segments, the lengths of each segment being equal to the eight segments of the A/H1N1 influenza virus. Five biallelic variant loci were introduced at random locations in each segment, with each loci having randomly drawn alleles. Combining the variant alleles in an exhaustive manner resulted in a total of 2^5 potential full haplotypes of which eight were chosen to represent the viral population. Facilitating comparisons across multiple simulations, each variant site was required to remain polymorphic in the chosen population, this being guaranteed by repeated sampling of the eight haplotypes until the criteria was met. Subsequently, full haplotype frequencies were generated at random, constrained by a minimum haplotype frequency of 5%, which was ensured through the repeated sampling of frequencies.

Transmission was modelled as a multinomial draw from the donor population with sampling depth equal to the bottleneck size. Selection for transmission was incorporated as a shift in haplotype frequencies as described in Equation 3.15. Where included in the simulation, selection was assumed to act upon a single variant in one of the viral segments with selection occurring prior to transmission. Within-host growth was modelled as a single round of replication defined

as a multinomial draw from the founder population conferring a 22-fold increase in population size.

Partial haplotype observations were generated on the basis of short-read data simulated for each gene. Short-reads were modelled as randomly placed gapped reads with mean read and gap lengths derived from an example influenza dataset (Wilker et al. 2013) (mean read length = 119.68, SD read length = 136.88, mean gap length = 61.96, SD gap length = 104.48, total read depth = 102825); these estimates are conservative relative to what can be achieved with the best contemporary sequencing technologies. Read depths were calculated for all possible sets of partial haplotypes by assigning individual reads to sets according to the loci they cover. Based on these read depths and the post-transmission viral population, a set of full haplotype observations were obtained using a Dirichlet-multinomial sampling process employing a dispersion parameter C to account for noise. Finally, partial haplotype observations were derived by summing over contributions from each of the associated full haplotypes.

Replicate experiments were generated by considering replicate observations of transmission events with consistent viral populations; that is, for which the variant alleles were consistent between replicate transmission events.

3.2.13 Data Processing Within Transmission Scheme

Simulation of transmission events led to the production of multi-locus variant data in the form of partial haplotypes. These were removed from consideration if A) the partial haplotype did not have at least 10 observations either before or after transmission, B) the partial haplotype exhibited a frequency of $< 1\%$ before transmission, C) the partial haplotype had no observations before transmission (variant assumed to have arisen *de novo*), D) the partial haplotype was the only partial haplotype in its set and had no observations post-transmission. Additionally, to avoid potential dataset errors from drastically influencing the inference outcome, partial haplotypes were removed if found to have a single post-transmission observation despite the presence of a large ($\geq$ 50) overall sampling depth. Finally, removal of partial haplotype observations may result in individual loci becoming monomorphic (all partial haplotypes covering these loci exhibit the same alleles). In this case, relevant partial haplotype sets were removed with the reads being redistributed unto variant sets with fewer loci.

SAMFIRE (Illingworth 2016) was used to construct a set of haplotypes spanning each viral segment using the multi-locus variant calls from all time points.

Here, potential haplotypes are identified by a process of exclusion. Given n biallelic variants in a segment, there are 2^n potential haplotypes, or combinations of those variants across all loci. SAMFIRE uses observed partial haplotype reads to constrain this set. For example, if across four loci only three of a potential sixteen combinations of alleles are observed, this removes a large proportion of the potential haplotype set. The haplotypes identified in this manner comprise the space of haplotypes spanned by the vectors $\boldsymbol{q}^B$ and $\boldsymbol{q}^A$. No inference of haplotype frequencies is conducted at this point, such inference is conducted in a subsequent step, using the likelihood framework described above.

3.2.14 Inference of Parameters

3.2.14.1 Hierarchical Selection Model

In our model, the set of potential fitness parameters is large. To simplify the calculation, parameters modelling three- or higher-locus epistatic effects were neglected, while parameters modelling two-locus epistasis were only considered for addition to a model which already contained single-locus fitness parameters for each of the two loci. In both the inferences of within-host selection and of transmissibility, a null assumption of neutrality was used as the starting point for an inference model, exploring successively more complex models of selection until an optimal model, defined according to a model selection process, was identified.

3.2.14.2 Replicate Calculations of Transmission Parameters

Both our within-host and transmission calculations are performed in a model space of potential haplotypes. For example, in the first step of the transmission model, I calculate an estimate for the population $\boldsymbol{q}^B$ given the data $\boldsymbol{x}^B$. In many cases, particularly where there are greater numbers of potential haplotypes and short reads span smaller numbers of loci, it is possible that the data $\boldsymbol{x}^B$ will not uniquely specify the initial vector $\boldsymbol{q}^B$ (see also Chapter 2). Here we are concerned about inferring parameters of transmission, rather than the explicit haplotype reconstruction. However, to check the robustness of the inference, statistical replicate calculations were run, using different reconstructions of $\boldsymbol{q}^B$ in each case; median inferred parameters across replicates are presented in the following. To improve the speed of the inference, haplotypes in $\boldsymbol{q}^B$ with inferred frequencies of less than 10^{-10} were removed from the calculation; subsequent to

this, haplotypes were removed in increasing size of inferred frequency until no more than 100 haplotypes remained in $\boldsymbol{q}^B$ at non-zero frequencies.

I note that the inference of $\boldsymbol{q}^B$ depends upon the initial identification of a plausible set of underlying viral haplotypes using SAMFIRE. A broad set of haplotypes is required for the comparison of different hypotheses about selection in the system. However, where the initial set of haplotypes is very large, as might occur where very short reads describe a great number of polymorphic loci, this approach to haplotype reconstruction becomes computationally challenging.

3.2.14.3 Model Selection

Model selection was performed using the Bayesian Information Criterion (Schwarz 1978):

$$\mathrm{BIC} = -2 \log L + K \log n \tag{3.48}$$

where L is the maximum likelihood obtained for a model, K is the number of parameters in the fitness model, and n is the number of data points. A range of potential fitness models were explored, the optimal model being identified as that for which the addition of any single fitness parameter failed to bring a significant improvement in BIC.

3.2.15 Adaptive BIC

Noting previous discussion of the complexity of using BIC in biological modelling (Fischer et al. 2014), I here adopted a machine-learning approach to the interpretation of BIC statistics. Classically, a difference of 10 units of BIC has been held to represent strong evidence in favour of the additional parameter (Kass and Raftery 1995). Consistent with previous approaches this heuristic was used in the inference of within-host selection; in this inference the final model parameters make only small refinements to the inferred fitness landscape (Illingworth 2015). In the inference of transmission, the default approach is insufficient as successive nested models may differ substantially in the size of the bottleneck they report. As such, using a fixed difference of 10 BIC units for model selection resulted in an overestimation of the extent of selection with a high false positive rate (see Section 3.3.3 and Figure 3.10). In order to generate a more robust approach to model selection, I generated and analysed simulated data to identify the optimal interpretation of BIC differences. Given a real dataset for analysis,

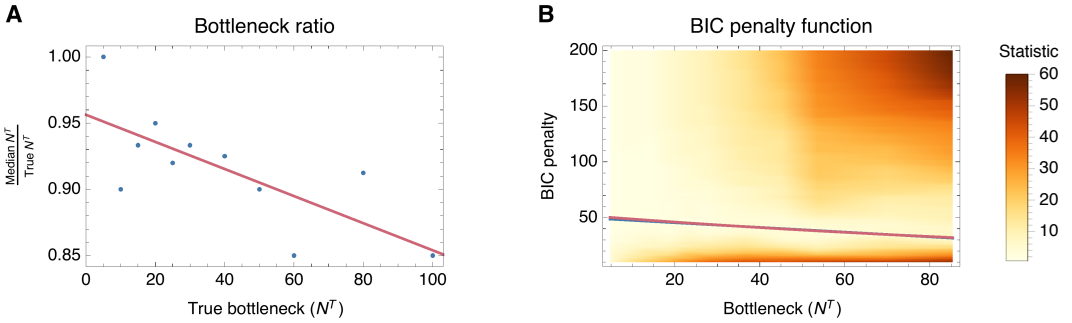


Figure 3.6. Determining BIC penalty function for bottleneck inference under simulated data. **A** The ratio of the median inferred bottleneck to the true bottleneck is plotted against the true bottleneck size. As shown in Figure 3.7, as the bottleneck increases, our ability to infer it correctly decreases due to noise. In order to account for this phenomenon, a straight line is fitted to the data aiming to capture the general trend. **B** Heat map of the bottleneck-specific statistic plotted against BIC penalty and bottleneck size. The plot was generated based on three datasets with selection coefficients $s = \{0, 1, 2\}$ and a simple statistic based on bottleneck differences was employed. More specifically, the median bottleneck was computed across 200 seeds and the bottleneck-statistic was defined as the absolute value of the difference between the median inferred bottleneck and the true bottleneck multiplied by the baseline determined in A). By considering bottlenecks in the range $[5, 100]$ and BIC penalty values in the range $[10, 200]$, a heat map was produced and linear and decay exponential regression were conducted seeking to minimise the sum of the statistic across the values of N^T that were considered.

simulated data was generated describing systems with equivalent numbers of gene segments and polymorphic loci to the real dataset, being observed with an equal number of reads spanning each set of loci, and with reads containing an amount of information specified by the parameter C inferred for the real dataset.

Next, inferences were conducted on data describing neutral transmission events with bottlenecks in the range $[5, 100]$. The ability to infer a correct neutral bottleneck is impaired by noise for transmission events involving a large number of viruses (see Section 3.3.1 and Figure 3.7). To correctly account for this, linear regression was used to obtain a simple function describing the ratio between the median inferred and true bottleneck sizes under neutrality (Figure 3.6A); this parameterises our expectation of the 'correct' inferred bottleneck size for any given real bottleneck, once noise is accounted for.

Secondly, using this baseline to set our expectations, a parameterisation was carried out to find a BIC penalty function that gave the largest accuracy in bottleneck inference. To this end, three datasets were considered; a neutral dataset and two datasets with single selection coefficients of $s = \{1, 2\}$ respectively. BIC penalty values in the range $[10, 200]$ were examined, with smaller BIC penalty

values leading to inferences with a larger number of selection coefficients and vice versa. For each BIC penalty value, the difference between the bottleneck inference of the optimal model (under BIC) and the baseline expectation was summed for the three datasets to give a statistic describing the accuracy of the inferred bottlenecks, this statistic being expressed as a function of the real transmission bottleneck N^T and the BIC penalty (Figure 3.6B). Finally, linear and decay exponential models were fitted to this data via regression, selecting the BIC penalty model which minimised the error in the inferred bottlenecks from the simulation data. We note that our penalty is a function of the inferred population bottleneck, higher penalties being inferred for tight bottlenecks and lower penalties being inferred for looser bottlenecks.

Thirdly, the inferred data were reinterpreted to derive a BIC penalty optimal for the inference of selection. Even with a BIC penalty function optimised for bottleneck inference, there may still remain cases where, through the stochastic process of transmission, the genetic composition of the population changes in a manner consistent with the action of selection, granting a false positive inference. A second BIC penalty was learned as above, this time maximising the accuracy of the inference or non-inference of selection parameters, defined as

$$\frac{\#\text{ true positives} + \#\text{ true negatives}}{\#\text{ true positives} + \#\text{ false positives} + \#\text{ true negatives} + \#\text{ false negatives}} \tag{3.49}$$

This conservative BIC penalty function typically led to an underestimate for the inferred bottleneck; the two BIC penalty functions were used in concert to estimate N^T and S^T in separate calculations. The BIC penalty functions are specific to individual datasets and, as a result, recalculation of BIC penalty functions is required when considering new data. Inference of BIC penalty functions is only necessary when attempting to jointly determine bottleneck and selection; for inference of transmission bottlenecks only the method is remarkably simple and fast.

As noted elsewhere, where a genomic variant fixes between two observations, this change in frequency can be explained by the fitting of an arbitrarily large selection coefficient; no upper bound on selection can be established (Illingworth and Mustonen 2012a). Within our framework, if this is not accounted for, extremely strong selection may be falsely inferred to explain the loss of variants during a transmission bottleneck. To guard against this, models of transmission in which the inferred magnitude of selection was outside of the range (-10,10)

were excluded from consideration. In the within-host analysis methods, haplotype fitness are not constrained; here, to avoid errors of machine precision, the magnitudes of extreme fitness inferences were reduced to be within the range (-10,10). For the same reason, elements of the mean and covariance matrix of $\boldsymbol{q}^B$ were constrained to be greater in magnitude than 10^{-11}. While selection coefficients outside of this range have been identified (Bull, Badgett and Wichman 2000), these steps greatly reduce the number of false inferences of strong selection.

3.2.16 Analysis of Simulated Data

I validated my model through the analysis of simulated data mimicking influenza transmission events. Potentially interesting scenarios were probed, e.g. varying bottleneck size, selective pressures and the extent of sequencing noise. I here briefly outline the simulation setup.

3.2.16.1 Accounting for Noise and Uncertainty

I initially analysed simulated data with a view to understanding the impact of sequencing noise and other sources of uncertainty. I considered bottleneck inference for scenarios wherein a range of bottlenecks were probed ($N^T = \{5, 10, 15, 20, 25, 30, 40, 50, 60, 80, 100\}$) and for which the extent of noise was varied ($C = \{50, 100, 200, 500, 10^3, 10^4, 10^5, 10^6\}$). I next investigated the accuracy of bottleneck inference given an incorrect estimate of the amount of sequencing noise present. Finally I considered the result of neglecting the variance component in $\boldsymbol{q}^B$.

3.2.16.2 Inference of Bottleneck Sizes and Selection for Transmission

Accuracy in bottleneck inference was tested both under the neutral (see Section 3.2.4.5 and Appendix C) and selective models. Bottlenecks were probed both in the presence and absence of selection. Where present, selection was defined to act on the third of five loci in HA with variable strengths, $\sigma^T = \{0.5, 0.75, 1.0, 2.0\}$. True and false positive rates of selection inference were also investigated. True positives were defined as inferences for which selection was inferred for the selected locus in the system. False positives were defined as inferences for which any neutral locus was predicted to be under selection. Selection inferences were carried out both under a model of adaptive BIC (see

Section 3.2.15) and one with a flat BIC penalty of 10 log likelihood units. Aiming to determine magnitudes of selection inferences, smooth kernel distributions were computed using the true positive outcomes.

3.2.16.3 Benchmarking Against Sobel Leonard et al. Method

The beta-binomial method of Sobel Leonard et al. was implemented as described in Section 3.2.11. The neutral version of our model was compared to the Sobel Leonard et al. method on the basis of simulated data. For this purpose neutral simulated data were generated with $C = 10^6$, i.e. noise-less sequencing, and growth factors of $g = \{1, 22\}$.

3.2.17 Online Repositories

Code and scripts related to transmission inference can be found online at `https://bitbucket.org/casperlu/transmission_project/`. The SAMFIRE (Illingworth 2015) data processing suite can be found at `https://github.com/cjri/samfire`. Detailed descriptions of code options and user guides are available in the repository README files. Scripts and instructions relevant for generating figures may be found online as well.

3.3 Results

This section considers the results obtained from applying the basic inference model to simulated data and from benchmarking against a current state-of-the-art inference framework.

3.3.1 Sequencing Noise Limits the Maximum Inferrable Bottleneck

Application of our model to simulated data describing neutral population bottlenecks showed that a lack of sequencing noise is critical for the correct inference of large population bottlenecks (Figure 3.7). Inferences of bottleneck sizes showed a limit on the inferred bottleneck size governed by noise in sequencing; where there was little noise in the data (i.e. at values of C greatly in excess of the bottleneck), a correct inference of the true population bottleneck was generally made. However, as noise increases, the inferred bottleneck reaches a plateau

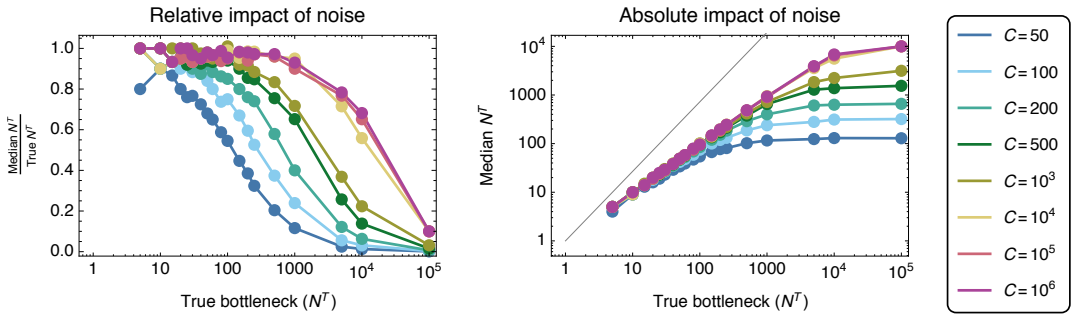


Figure 3.7. Influence of sequencing noise upon the ability to infer a population bottleneck size from genome sequence data. Median inferred bottlenecks are shown, calculated on the basis of 200 replicate simulations for each point. In the left-hand plot, a value of 1 indicates a correct bottleneck inference; in the right-hand plot, the absolute inferred bottleneck size is shown. Simulations were conducted under the assumption of selective neutrality, with no attempt to infer selection from the data.

above which increases in the true bottleneck no longer affect the inferred bottleneck size. This result can be understood in terms of the extent to which the population bottleneck and noise contribute to the change in the viral population; where large numbers of viruses are transmitted, most of this signal is likely to result from noise. Here we note failures in the inferred bottleneck size even with very high C; these occur due to the finite read depth in our simulations, which was of order 10^4. In these calculations a neutral method, in which selection was assumed to have no effect on the population, was used to make inferences from neutral simulations. A consistent value of C was used for simulation and inference purposes.

In a real dataset the extent of noise may be unknown. Further investigation showed bottleneck estimation to be relatively robust to an incorrect estimate of the extent of noise in a dataset, except where the extent of noise was substantially overestimated (Figure 3.8). In general, an underestimate of the extent of noise in a dataset led to an inferred bottleneck size that was marginally lower than the value obtained given the true amount of noise; for example where the value $C = 10^6$ was used to infer a bottleneck from data with $C = 50$, a bottleneck of true size $N^T = 50$ was inferred as $N^T = 25$. An overestimate of the extent of noise led to an overestimate of the size of the bottleneck with severe overestimation resulting in dramatically incorrect inferences. Therefore, while noise limits the potential of a method to identify large bottleneck sizes, underestimating the extent of noise in the data is generally the safer approach.

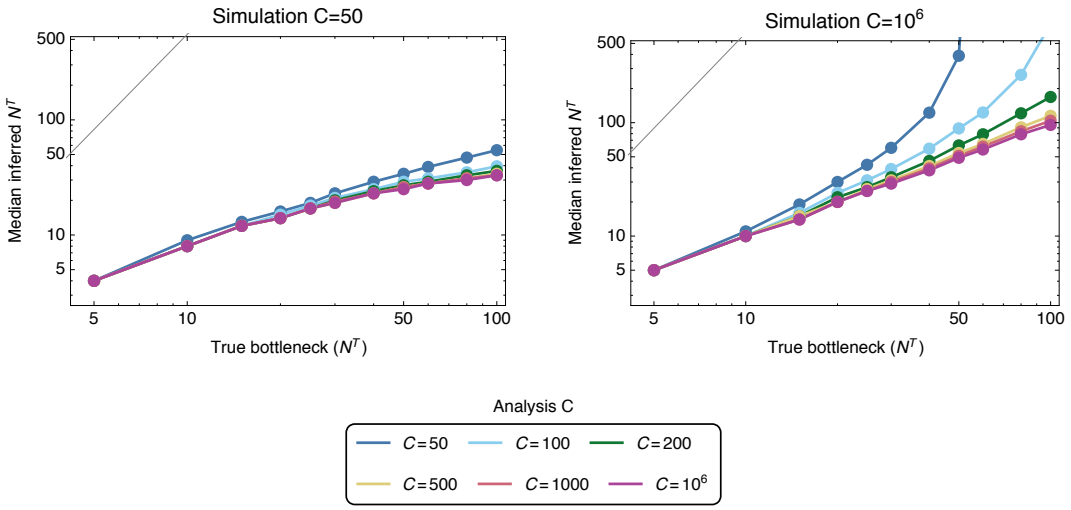


Figure 3.8. Bottleneck inference under a neutral model applied to neutral data with simulation dispersion parameters of $C = \{50, 10^6\}$. Inference was performed using a range of dispersion parameters, $C = \{50, 100, 200, 500, 1000, 10^6\}$. Each datapoint represents a median over 200 simulation seeds.

3.3.2 Accounting for Uncertainty in $\boldsymbol{q}^B$

Within the transmission framework we represent the pre-transmission viral population by a multivariate normal distribution in $\boldsymbol{\mu}^B$ and Σ^B. The matrix Σ^B is challenging to specify in full due to the limited amount of information available; we therefore define its non-diagonal entries to be zero, i.e. we ignore between-haplotype contributions to the uncertainty in $\boldsymbol{\mu}^B$. An alternative parameterisation strategy is to consider $\boldsymbol{q}^B$ as a point vector, i.e. $\Sigma^B = 0$. Despite the mathematical convenience and associated efficiency increase, this approach resulted in considerable underestimation of bottleneck sizes as shown in Figure 3.9. This result highlights the need for sufficient account to be taken of the level of uncertainty in variables determined from noisy data.

3.3.3 BIC Considerations

The Bayesian Information Criterion (BIC) allows for comparison of models of differing complexity. A model of increasing complexity bringing about an improvement in BIC of more than 10 log likelihood units have traditionally been considered as strong evidence in favour of this more complicated model (Kass and Raftery 1995). Attempts at inferring selection in transmission events under a fixed BIC penalty of 10 units are shown in Figure 3.10. A considerable amount of overfitting was found as evident from a simultaneously large true and

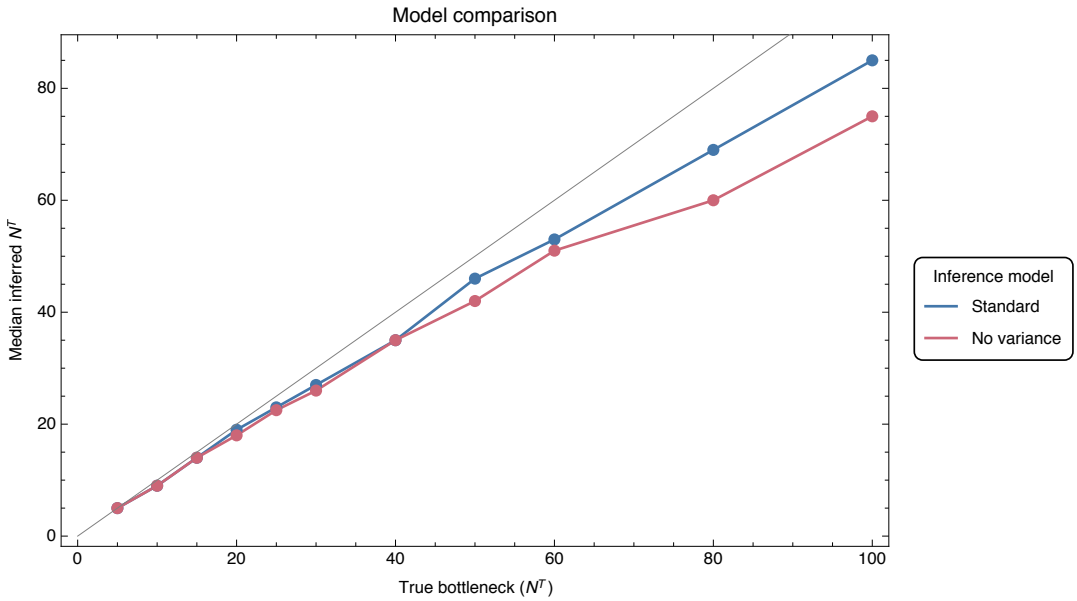


Figure 3.9. Median inferred bottleneck size from simulated neutral transmission data. Inferences were made using either the standard neutral model, in which the covariance matrix $\boldsymbol{q}^B$ is diagonal, or using a simplified model ignoring the variance in $\boldsymbol{q}^B$ altogether. Each datapoint represents a median over 200 simulation seeds.

false positive rate of selection inference. Additionally, as the true bottleneck increases, the false positive rate decreases whilst the true positive rate improves.

This overfitting arises from the nature of the problem; given only two datapoints it is possible to arbitrarily assign selection to perfectly explain the change in the viral population across the transmission event. If selection is allowed to explain every aspect of the diversity change, the method would effectively infer an infinite bottleneck. Even worse, in the case of extinction events, infinitely large selection coefficients may be inferred. We guard against this by limiting the magnitude of selection inferences to ± 10 and discarding selection models with inferences in excess of ± 9, see Section 3.2.15.

Additionally, the standard BIC framework is unable to properly account for the amount of information contained within a biological dataset. For instance, in the case of genomic data, two datasets may have the same read depth but exhibit wildly different read lengths. Naturally, the dataset with longer read lengths is more informative, but a BIC framework based around read depth alone doesn't capture this. Similarly, within our framework, partial haplotype datasets are weighted evenly (see Equation 3.37) despite the fact that a partial haplotype set covering 5 loci are considerably more informative than one covering a single variant. In other words, a sequence read does not represent a standardised

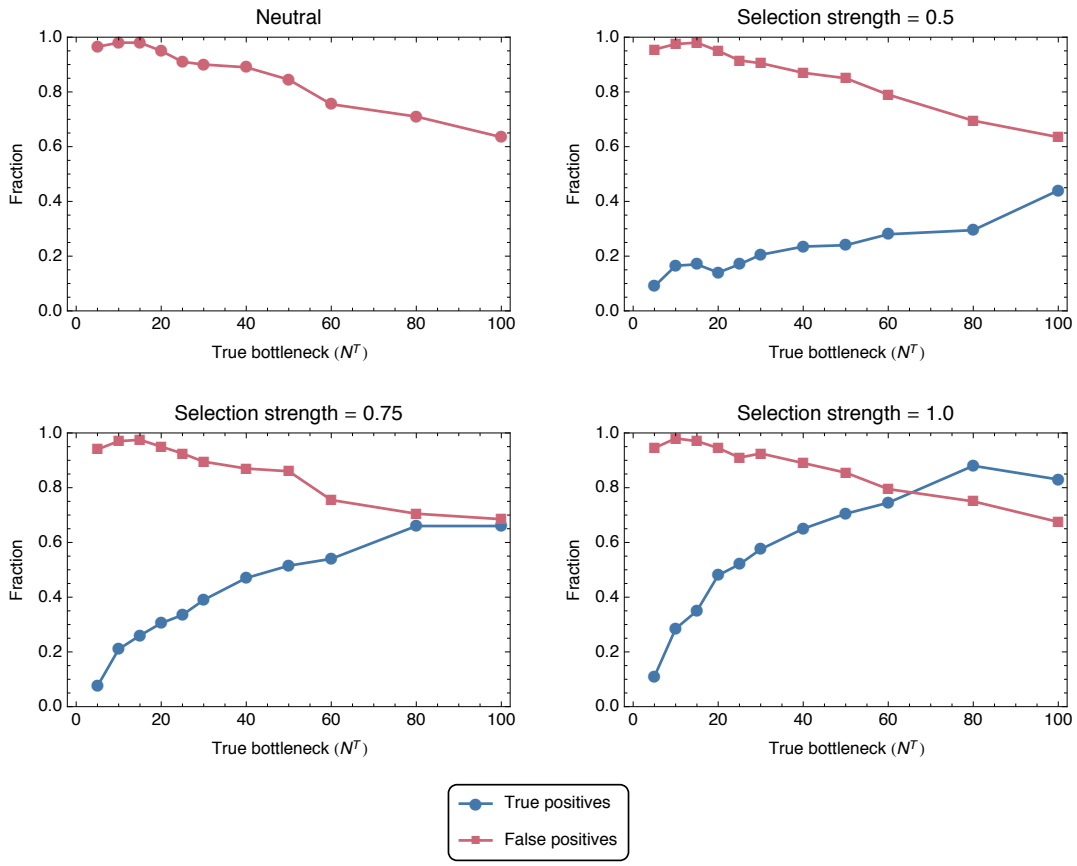


Figure 3.10. True and false positive rates of selection inference from 200 simulations of transmission events from single-replicate systems in which a single variant was under selective pressure for increased transmissibility of $\sigma \in \{0, 0.5, 0.75.1.0\}$. True positives were defined as inferences for which selection was inferred for the selected locus in a system; false positives were defined as inferences for which selection was inferred at any neutral locus or for multiple neutral loci in the system. Inferences can be simultaneously true and false positive, i.e. the true and false positives rates are not required to sum to unity. A fixed BIC difference of 10 units were employed in the model selection process, requiring a model with a single additional parameter to generate an improvement of at least 10 units in BIC to be accepted. While such a difference is generally accepted as showing strong evidence in favour of the more complex model, in our case it generated a high rate of false positive inferences of selection.

unit of information. Lastly, the magnitude of the BIC improvement required to accept a model varies with bottleneck size. As the bottleneck size increases, stochastic changes diminish allowing for a larger accuracy in selection inference, i.e. a smaller BIC penalty is required. Taken together, we here propose the need for an 'adaptive BIC penalty', accounting on the one hand for the amount of available information and on the other for the impact changes in bottleneck size has upon selection inference.

3.3.4 Variance in Inferred Transmission Bottlenecks

Results from individual simulations showed that the method could discriminate between bottleneck sizes that differ by a factor of three or above (Figure 3.11 and Figure 3.14, top left plot). Obtaining precision in an estimated bottleneck or effective population size is inherently a difficult task, relying on the estimate of the extent of a stochastic effect from limited data (Malaspinas et al. 2012). Across 200 simulations, the interquartile range in an inferred bottleneck spanned close to 28% of the true bottleneck size, with inferred values spanning a range of approximately 130% of the correct bottleneck size. A slight underestimate in the bottleneck size for the case $N^T = 100$ was consistent with the extent of noise in sequencing; here and in all subsequent simulations a value of $C = 200$ was used, representing an extent of noise that is readily achievable from short read sequence data (Illingworth 2015; Illingworth et al. 2017). In our inferences, while gross differences in bottleneck size can be identified, a high level of precision is difficult to obtain from sequence data alone.

3.3.5 Inference of Population Bottleneck Sizes Under Selection for Transmission

Inferences of bottleneck size showed a systematic underestimation of the bottleneck when selection affected a transmission event, but a method neglecting selection was used in the inference procedure (Figure 3.12). Simulations were conducted in which an allele at the third of five polymorphic loci in the HA segment of a simulated influenza virus increased the transmissibility of the virus according to a selection coefficient σ; this model of selection was applied for all subsequent simulations. In our simulations a value of $\sigma = 1$ is equivalent to a change in the frequency of a variant from 50% to 73% in a single transmission

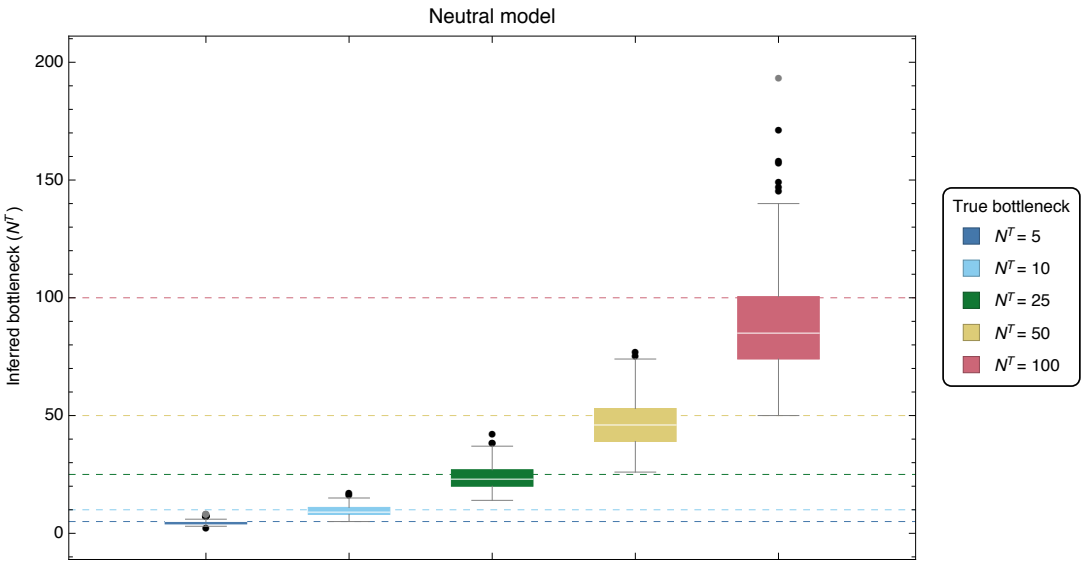


Figure 3.11. Inferred bottleneck sizes (N^T) for true bottlenecks $N^T = \{5, 10, 25, 50, 100\}$. Results were generated by applying a neutral inference model to neutral simulated data. Inferences are shown for 200 simulations at each bottleneck size.

event. The relatively strong magnitudes of selection considered reflect the short period of time (a single generation) over which selection for increased transmissibility can act and the relatively small number of viruses likely to be involved in a transmission event.

Inferences of population bottleneck were conducted using a neutral inference method, and with a model in which selection was not constrained to be zero. In the first case, ignoring selection led to an underestimation of the true bottleneck size by an amount which increased according to the magnitude of selection for transmissibility. Selection during transmission produces a shift in the expected composition of the viral population; if this shift is interpreted as occurring solely due to a finite bottleneck, a tighter bottleneck, inducing a larger stochastic change in the population, is inferred. This understanding explains the more pronounced underestimates achieved at larger bottleneck sizes; larger bottlenecks produce smaller stochastic changes in the population relative to the change induced by selection. When the full version of our model was run, allowing for a consideration of selection effects, the median bottleneck inferred from data under selection resembled that inferred from neutral data; the small shortfalls in the inference from neutral data are here explained by the influence of noise.

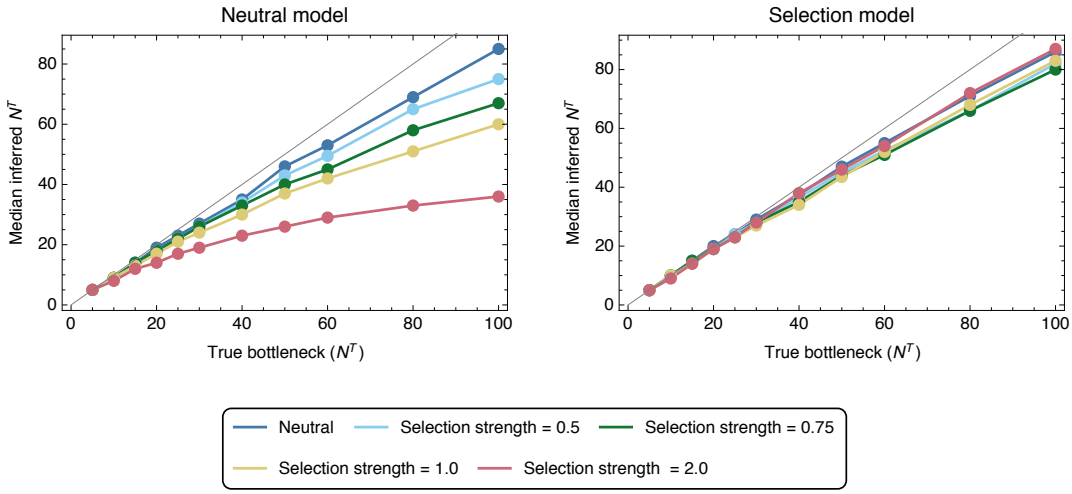


Figure 3.12. Median inferred bottleneck size from data simulating transmission with a single locus under selection of magnitude $\sigma \in \{0, 0.5, 0.75, 1.0, 2.0\}$. Inferences were made using either a neutral inference model, in which the effect of selection was assumed to be zero, or a model incorporating selection, which allowed the presence of selection to be inferred. Median inferences are shown from 200 simulations for each data point.

Calculations performed for data describing multiple replicate transmission events gave similar inferred transmission bottlenecks to those obtained from single replicates. In each case sets of three replicate transmission events were simulated, each event involving the transmission of virus between a distinct pair of hosts. Simulating the use of a consistent inoculum, our transmitted populations shared a common set of polymorphic loci in each segment. Median inferred values are shown in Figure 3.13. Full results describing the range of inferred bottleneck sizes from both one- and three-replicate populations are shown in Figures 3.14 to 3.17.

3.3.6 Identification of Variants Under Selection

In contrast to measures of diversity, which attempt to associate selection with a gene or segment of a virus, our method was able to correctly identify specific variants conferring increased transmissibility. Success was more often achieved in cases for which selection was relatively strong and the transmission bottleneck was relatively large (Figure 3.18). Our process for distinguishing selection from neutrality (Figure 3.5) can be tuned to identify a greater number of true variants under selection at the cost of making a greater number of false positive calls; here a conservative approach to identifying selection was applied. I eval-

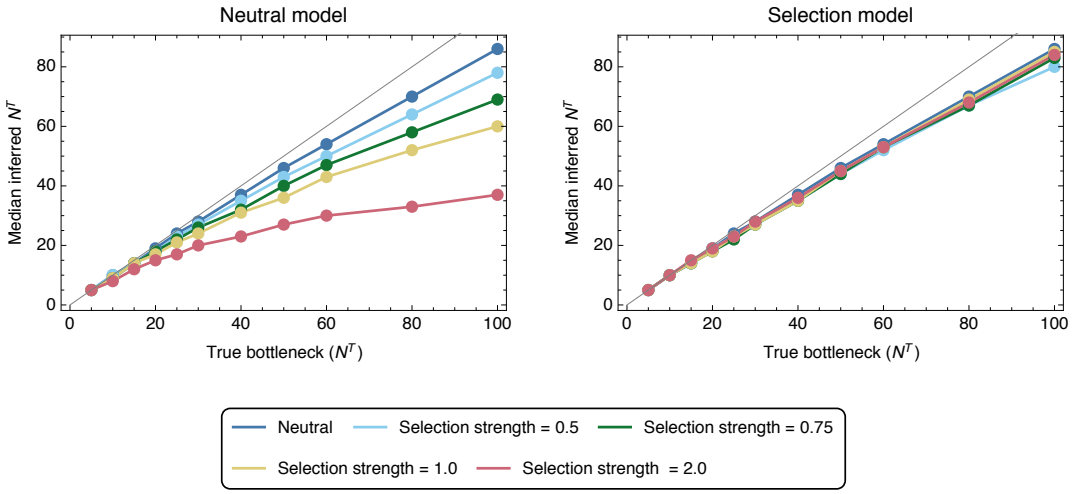


Figure 3.13. Median inferred bottleneck size from data simulating neutral transmission and transmission with a single locus under selection of magnitude $\sigma \in \{0, 0.5, 0.75, 1.0, 2.0\}$. Inferences were made using either a neutral model, in which the effect of selection was assumed to be zero, or a selection model, which allowed scenarios involving selection to be identified. Median inferences are shown from 200 simulations, each involving three replicate transmission events, for each datapoint.

uated populations in which a single locus was under selection, determining the potential to identify the variant. Under this approach the method retained a false positive rate (inference of selection at a non-selected locus) of 8% or less across the systems tested. Where a single variant was under a lower magnitude of selection ($\sigma \leq 0.5$), correctly identifying sites under selection was very difficult, though as selection became stronger ($\sigma \geq 1$) loci under selection could be identified with greater accuracy. Where selection existed the potential for it to be identified was greater at larger bottleneck sizes. These results can again be understood with respect to the dynamics of the system. The bottleneck has a stochastic effect on the population of a magnitude inversely related to the number of viruses transmitted. Inferring the presence of selection requires the identification of changes in the population going beyond what would be expected under neutrality, biasing the population in the direction of the selected allele or alleles. However, stochastic effects can by chance distort the population in one direction or another by more than the expectation; this leads to false inferences of selection. Genuine changes resulting from selection become easier to identify when the changes are themselves larger (stronger selection) or where the magnitude of the stochastic effect is reduced (higher N^T). While data from multiple replicate simulations made little difference to the inferred bottleneck

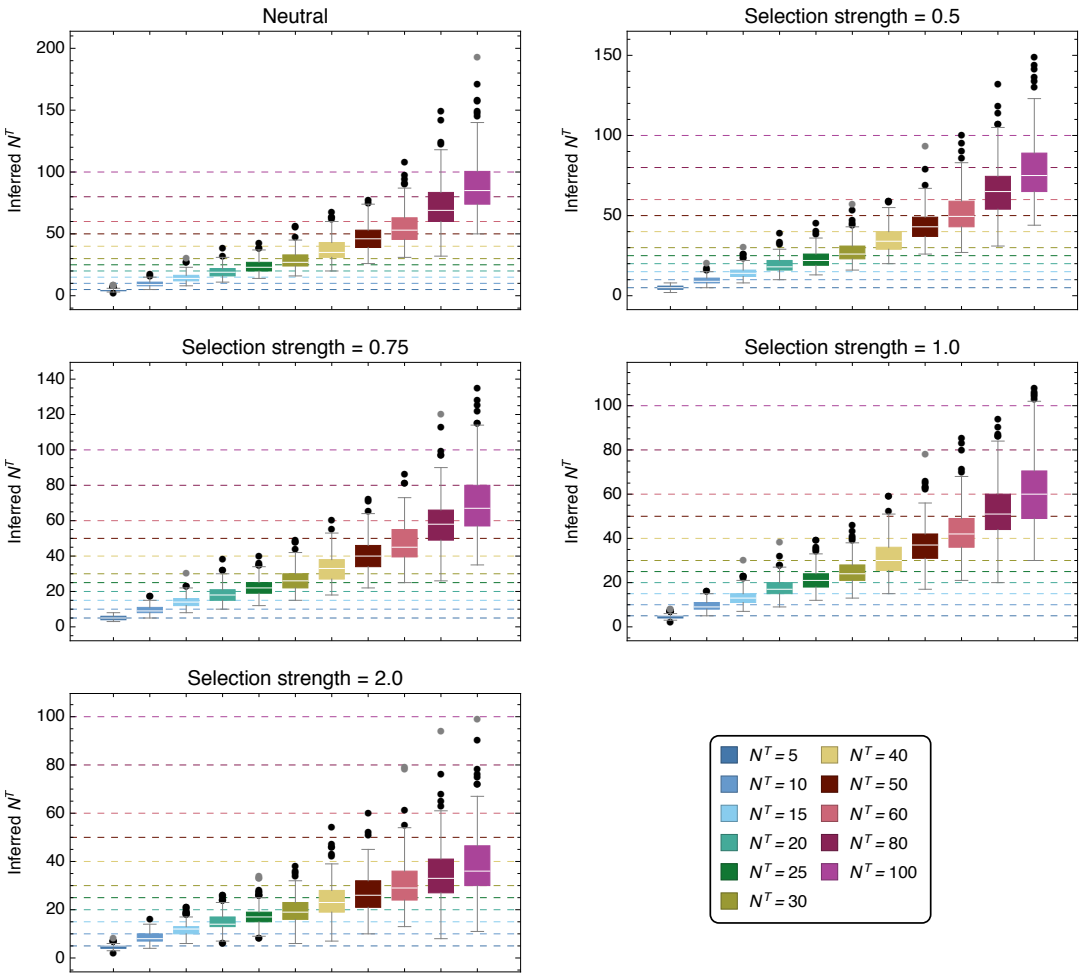


Figure 3.14. Inferred bottleneck sizes N^T for a range of true bottleneck sizes. Results were generated by applying a neutral inference model to selected simulated data. Results are shown for 200 simulations at each bottleneck size.

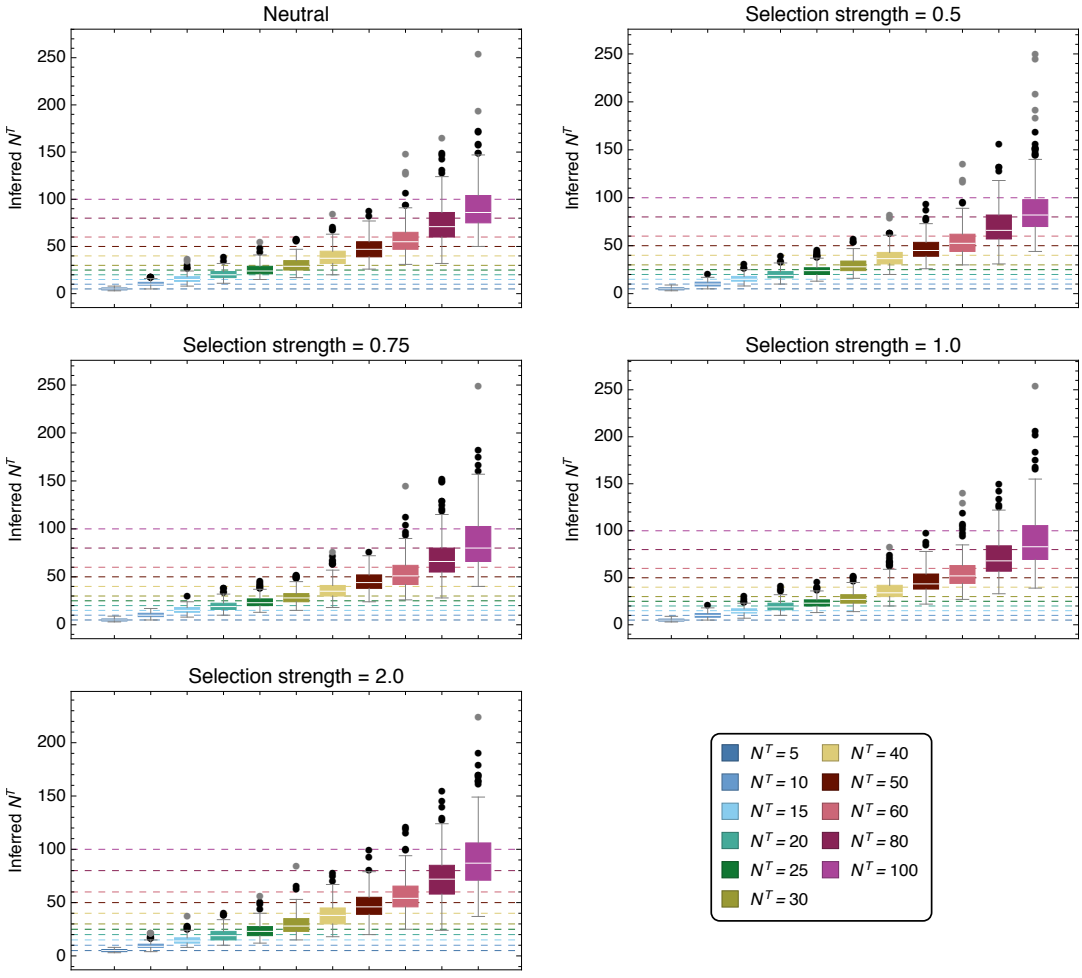


Figure 3.15. Inferred bottleneck sizes N^T for a range of true bottleneck sizes. Results were generated by applying an inference model accounting for selection to selected simulated data. Results are shown for 200 simulations at each bottleneck size.

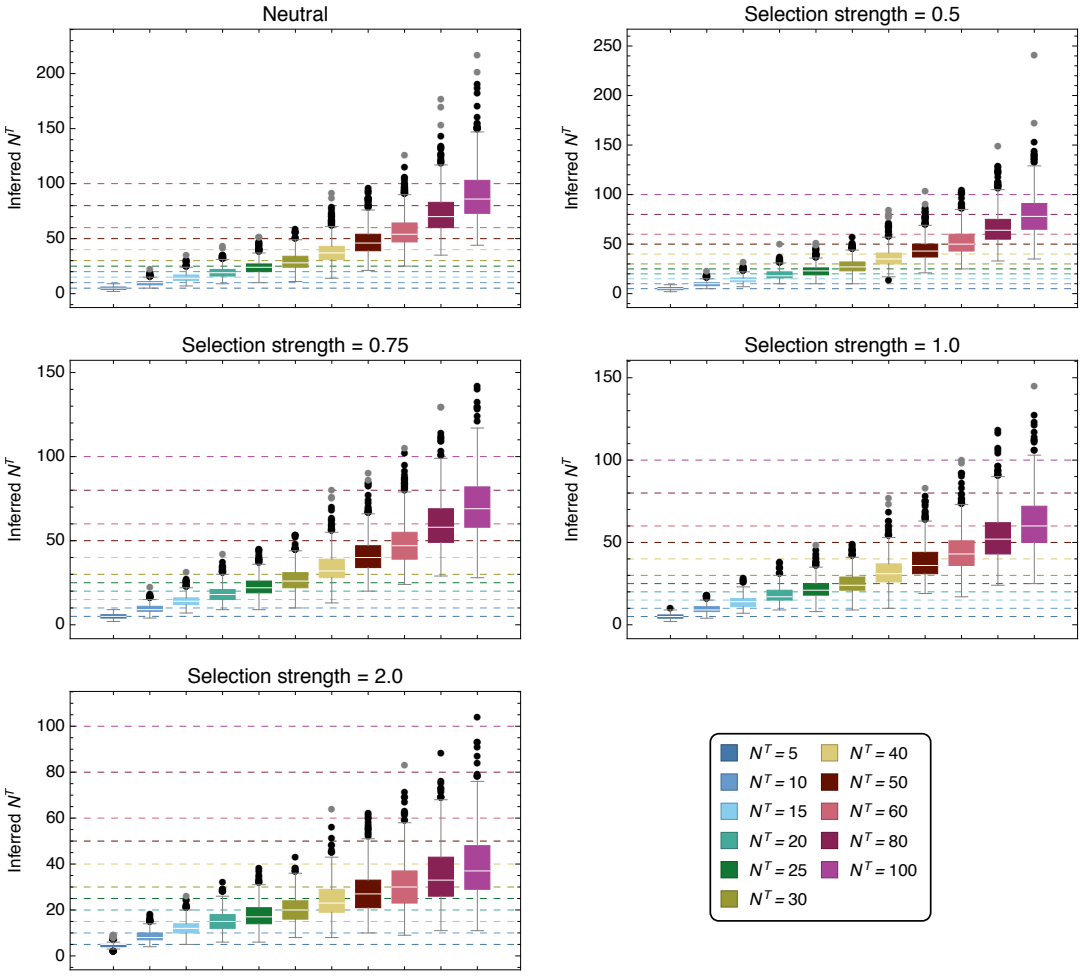


Figure 3.16. Inferred bottleneck sizes N^T for a range of true bottleneck sizes. Results were generated by applying a neutral inference model to selected simulated data. Results are shown for 200 simulations at each bottleneck size, each simulation describing three replicate transmission events.

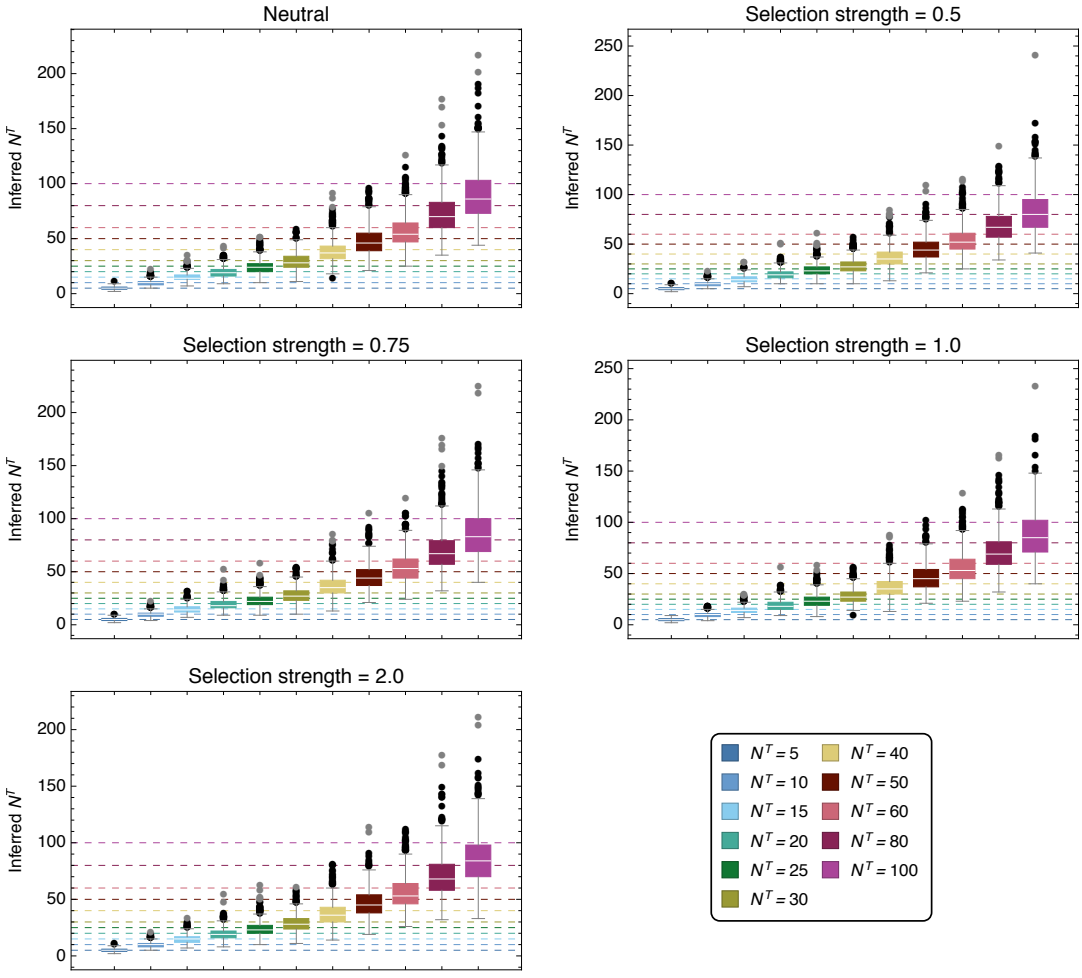


Figure 3.17. Inferred bottleneck sizes N^T for a range of true bottleneck sizes. Results were generated by applying an inference model accounting for selection to selected simulated data. Results are shown for 200 simulations at each bottleneck size, each simulation describing three replicate transmission events.

size (see above), such data led to a more dramatic change in these results, with the false positive rate falling to zero for bottlenecks with $N^T \geq 20$. The power of replicate experiments arises from the lower probability that stochastic effects will impose a consistent pattern of change upon multiple populations. While a larger-than-expected stochastic change in the frequency of a variant may occur in one system, leading to a false positive inference of selection, it is unlikely that the same pattern would recur across multiple replicates. While the inference of selection for transmissibility is not easy, the use of replicate experiments is of considerable value in this task; while, under our conservative approach, not all variants truly under selection were identified, those which were identified from replicate data were almost universally true positive calls.

3.3.7 Estimating the Magnitude of a Selected Variant

Given the correct identification of selection acting for a specific variant, the inferred magnitude of selection was marginally overestimated, with an increased overestimate at smaller values of the transmission bottleneck N^T (Figure 3.19). The mixture of deterministic and stochastic changes in the population explains this phenomenon; the population after transmission is equal to its expected value plus some stochastic change. In the event that the stochastic change is aligned with the direction of selection, the presence of selection is more likely to be inferred, while the additional change in that direction will give an overestimate of selection. Conversely, if the stochastic change is in a direction opposed to the influence of selection, the presence of selection is less likely to be inferred. Thus, selection was disproportionately inferred to exist when stochastic changes in the population led to an overestimate of its magnitude. Inferences conducted on sets of replicate transmission events produced more accurate and more precise estimates of selection. For example given a bottleneck of $N^T = 100$ and a true strength of selection of 0.75, the mean inferred selection from a single replicate was 1.00 with variance 0.040, while the mean inferred selection from three replicates was 0.90 with variance 0.010. (Figure 3.20)

3.3.8 The Biology of Within-Host Viral Growth May Affect the Inference of Transmission Bottlenecks

Comparing our approach with a previous inference method, we found that the biology underlying within-host viral growth can significantly affect the inferred

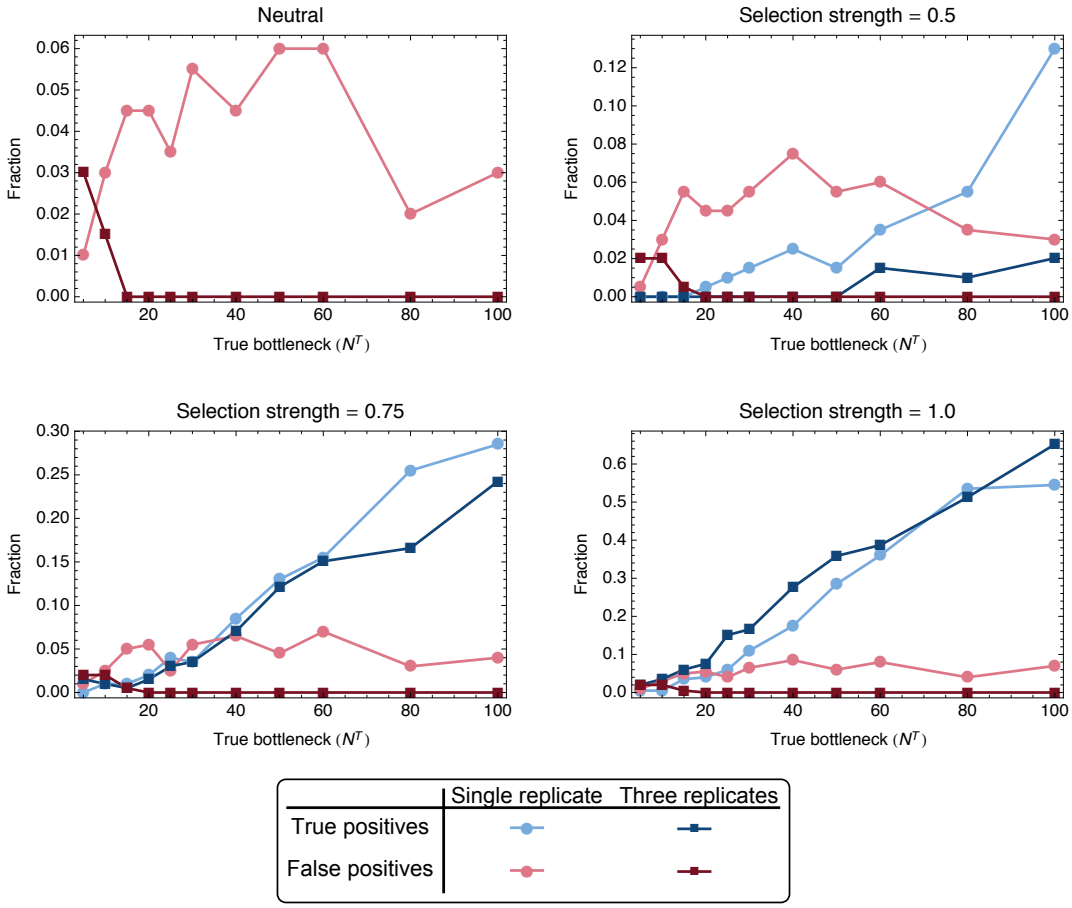


Figure 3.18. True and false positive rates of selection inference from 200 simulations of transmission events from single- and three-replicate systems in which a single variant was under selective pressure for increased transmissibility of $\sigma \in \{0, 0.5, 0.75.1.0\}$. True positives were defined as inferences for which selection was inferred for the selected locus in a system; false positives were defined as inferences for which selection was inferred at any neutral locus or for multiple neutral loci in the system. Inferences can be simultaneously true and false positive (see e.g. Figure 3.10).

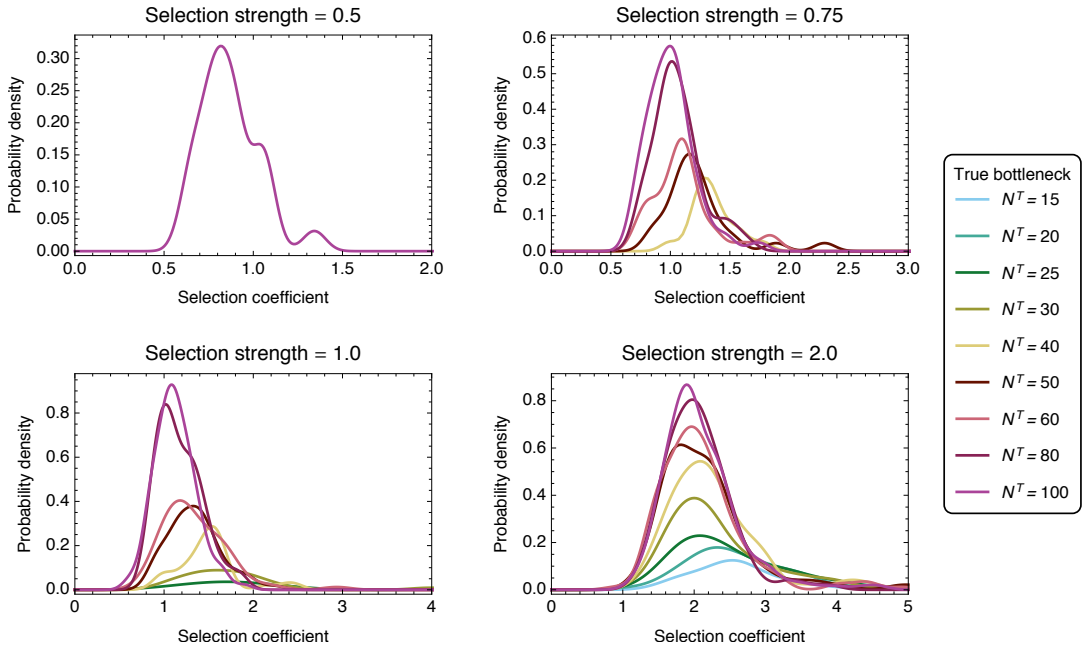


Figure 3.19. Probability distributions of inferred selection coefficients from 200 simulations of transmission events with selective pressures $\sigma \in \{0.5, 0.75, 1.0, 2.0\}$. Distributions were constructed for bottleneck values where the inference of selection resulted in a true positive rate for identifying selected variants of above 5 %. Smooth kernel distributions were computed using a Gaussian kernel function defined on $(0, 10)$ and Silverman's rule of thumb (Silverman 1986, p. 48) employed for the bandwidth size. Distributions were scaled such that their integral across the kernel range equalled the true positive rate.

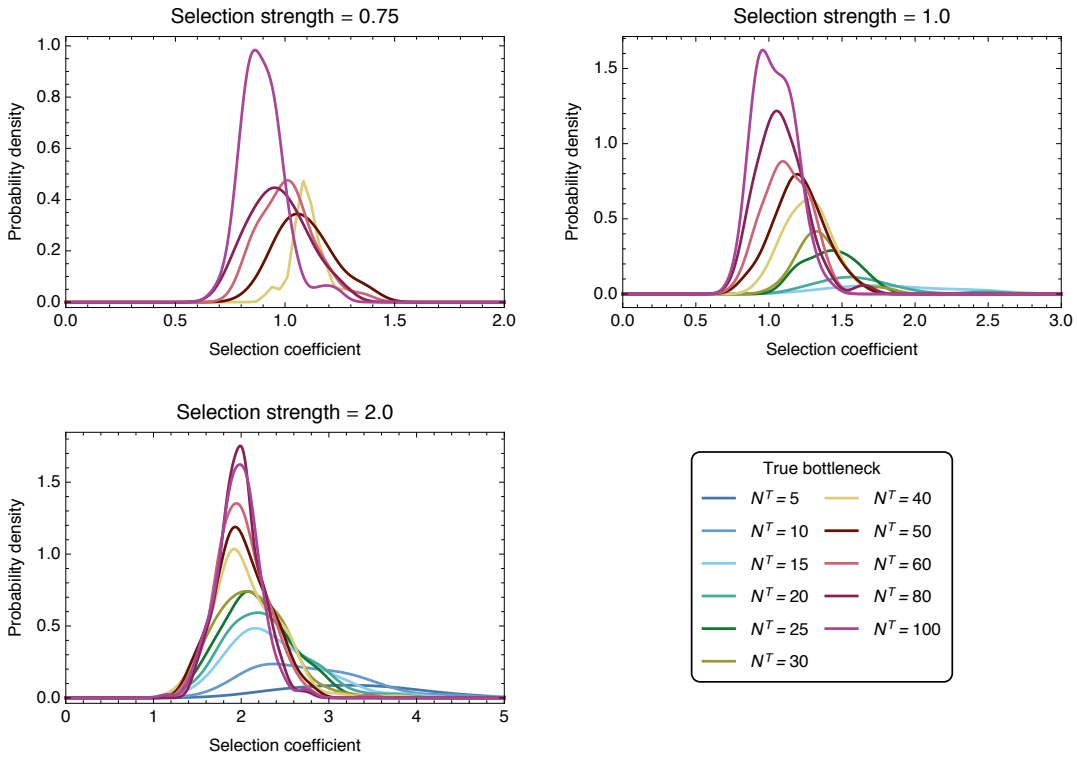


Figure 3.20. Probability distributions of inferred selection coefficients from 200 simulations each of three transmission events with selective pressures $\sigma \in \{0.75, 1.0, 2.0\}$. Distributions were constructed for bottleneck values where the inference of selection resulted in a true positive rate for identifying selected variants of above 5 %. Smooth kernel distributions were computed as for Figure 3.19

.

population bottleneck. In so far as previous population genetic models have not accounted for the presence of selection or noise in sequencing data (beyond binomial variance) I applied methods to data describing neutral transmission between a single pair of hosts under the assumption of error-free sequencing of samples. The method of Poon et al. (Poon et al. 2016) is explicitly defined across multiple transmission events so cannot be used to evaluate single transmission events. For this reason comparison was performed with the method described by Sobel Leonard et al. (Sobel Leonard et al. 2017b); I believe that this recent and well-cited approach, which infers a transmission bottleneck based on allele frequency change in a manner that accounts for within-host viral growth, represents the present state-of-the-art for bottleneck inference.

Comparison of the two methods showed our approach to have an increased flexibility to obtain correct inferences of population bottleneck size across a range of biological models of within-host growth. By default, our simulation model describes genetic drift during the within-host growth of the viral population as a single generation of replication, according to a Wright-Fisher population model with effective population size gN^T, where g is nominally the growth rate of the population; our inference framework was set to match the generative model (Figure 3.21). At a growth factor of 1, both methods correctly inferred the size of the population bottleneck. However, at our default growth factor of 22 (based upon experimental results in influenza (Baccam et al. 2006)), the method of Sobel Leonard et al., inferred a bottleneck size roughly double the correct value while the performance of our model was unchanged.

This result highlights the need to correctly account for within-host growth during the inference of a transmission bottleneck. If too much of the difference between the populations observed before and after transmission is accounted for by within-host genetic drift, the inferred bottleneck will be too high. By contrast, if not enough of this difference is accounted for as drift, the inferred bottleneck will be too low. In the approach of Sobel Leonard et al., the accounting made for genetic drift accounts for a variance equivalent to that incurred in a Wright-Fisher step of size N^T, that is, with $g = 1$ (personal correspondence, Daniel Weissman), obtaining a correct inference under these circumstances.

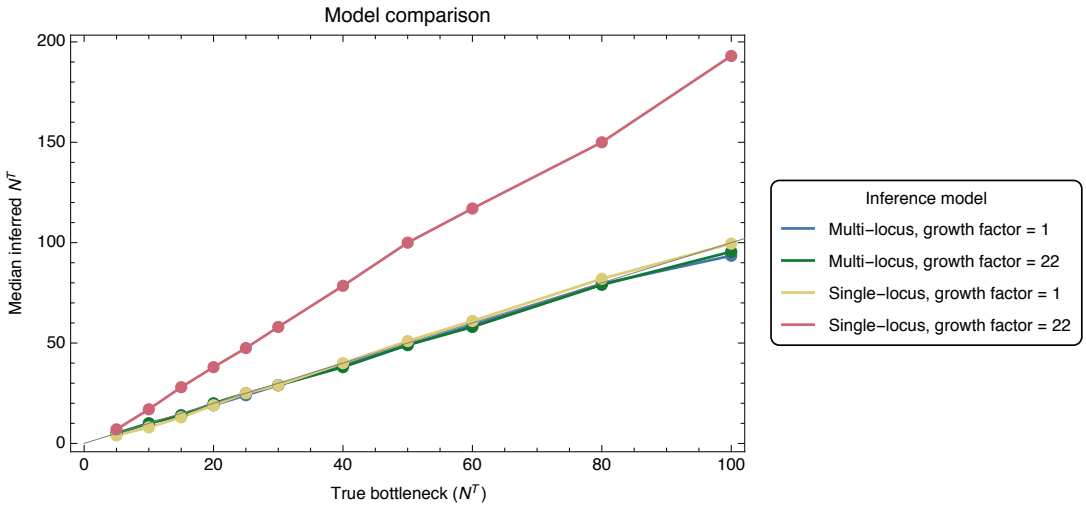


Figure 3.21. Median inferred bottleneck size from data simulating neutral transmission with the viral population undergoing either a single- or 22-fold increase in population size during within-host replication. Inferences were made using our approach termed the multi-locus method, which allows for specifying different growth factors, and the method of Sobel Leonard et al. (Sobel Leonard et al. 2017b), termed the single-locus method. Each data point represents the median bottleneck calculated over 200 replicate simulations.

3.4 Discussion

I have here presented an approach for jointly inferring a population bottleneck size and selection for differential transmissibility from viral sequence data describing a transmission event. While basic sampling approaches to bottleneck inference have been improved by an accounting for drift during within-host viral growth (McCrone et al. 2018; Poon et al. 2016; Sacristan et al. 2003; Sobel Leonard et al. 2017a), our approach additionally accounts for noise in genome sequence data, exploits partial haplotype data available from short-read sequencing, and separates the influence of a finite bottleneck from that induced by selection for increased transmissibility. In multiple studies, the transmission bottleneck has been found to be narrow during natural viral spread between hosts (Zwart and Elena 2015). While acknowledging previous evidence for the existence of small transmission bottlenecks in viral systems, I here note that a failure to account for selection and noise in the transmission process can decrease the bottleneck that is inferred from sequence data. Our approach is suitable for the analysis of acute infectious diseases such as influenza on the basis of a small number of observed transmission events; I note that where more substantial diversity is present in a within-host viral population, or where data are available

from a large number of hosts in an outbreak, phylogenetic methods of evolutionary inference become of increasing value (Derdeyn et al. 2004; Edwards et al. 2006; L. M. Li, Grassly and Fraser 2017).

Our study shows that the identification of variants conferring increased viral transmissibility is difficult when the number of transmitted viral particles is small. While improvements to our method may be achievable, this difficulty is fundamentally rooted in the nature of a transmission event; where a low number of virions transmit, the influence of stochastic processes becomes large, with variants fixing during transmission in a manner that cannot be distinguished from a selective sweep. The potential to infer the presence of selection increases at larger population size and given a greater number of replicate transmission events. However the amount of data required to make a statistically robust identification of a variant increasing viral transmissibility may be large. I note that, unlike more general inferences of selection from changes in viral diversity, our approach evaluates selection in terms of specific variants conveying an advantage or disadvantage for transmission. Where broad measures of diversity are calculated across segments of a genome, the background of genetic diversity across a large number of positions may be hard to separate from changes at individual positions under the action of selection.

When it comes to modelling selection during transmission, this is made difficult by selection not representing a single, clearly defined process. Rather, selection may occur during all aspects of the transmission process; as the virus is cleaved from the host cell and expelled from the host respiratory system, as the virus is airborne, and as the virus infiltrates the recipient airways and attaches to a new cell. As such, it may not be intuitively clear as to how we formally differentiate effects of within-host selection from those of selection from transmission. For instance, one may take the view that selection for transmission solely encompasses selection acting on the virus outside the host environment, with all other adaptive pressures representing within-host selection. Within this understanding, the bottleneck, which is a non-selective reduction in population size, occurs as the virus enters the respiratory tract of the recipient host with any subsequent adaptation following as a result of within-host selection. An opposing view is that selection for transmission includes all adaptive processes not directly occurring as a result of within-host growth. In this work I take this view, being interested in accounting for all aspects of the transmission process resulting in increased transmissibility of a virus. To this end, I model selection as a single event occurring prior to the action of the transmission bottleneck, the

bottleneck itself representing the founder population, i.e. the number of viruses that successfully establish a new infection. In other words, if only a subset of viruses are fit enough to be transmitted, the bottleneck represents a multinomial process dictating which of these viruses ultimately make up the founder infection. I note here, that if the order of selection and bottleneck was reversed, this would limit our ability to infer selection as the bottleneck reduces the diversity upon which selection acts. This phenomenon will be discussed in greater detail in the next chapter. I here make the assumption that selection acts prior to the bottleneck, noting that, where present, selection is likely to be the primary reason as to why a virus is either transmitted or not.

Our study provides some insight into the potential for inferring transmissibility using small animal experiments. One approach to exploring transmissibility (in influenza virus) has been the comparison, for different viruses, of the proportion of distinct animal pairs between which transmission occurs (Yen et al. 2011). The statistical significance achievable in these studies is limited by the number of animal pairs that can be examined (Linster et al. 2014; Nishiura, Yen and Cowling 2013; Steel et al. 2009). Furthermore, the comparison between one genotype and another may be confounded by viral heterogeneity, whereby each population contains a cloud of genetic diversity (Dinis et al. 2016; Wilker et al. 2013). As I have shown, data from replicate transmission events lead to an improved ability to infer selection, in particular by reducing the false positive rate of inference and by increasing the accuracy in inferred selection coefficients. I note, however, that the number of viral particles transmitted in each event is key in determining whether increased transmissibility can be identified; where a transmission bottleneck is narrow it is inherently difficult to identify selection against a background of large changes in the population induced by stochastic effects. Where transmission bottlenecks are small, a large number of replicates might be needed to make statistically well-supported inferences of increased transmissibility. Applications of our method to simulated data could be used to gain an insight into what might be obtained from a particular experimental setup.

Another aspect that affects the precision of transmission inference is the amount of available diversity. While not explicitly explored here, we can generate an intuition for the impact of donor diversity upon inference. Naturally, a minimum amount of diversity in the donor individual is required in order to infer a transmission bottleneck; where no diversity is present we cannot make any predictions. Considering the minimum amount of diversity possible, i.e.

a single locus polymorphism, this variant site harbours the largest amount of information when at intermediate frequencies, i.e. close to 50%. The closer the frequency is to the boundaries, the less power we have to distinguish large changes in diversity (narrow bottlenecks or strong selection) from small changes (loose bottlenecks or weak selection) given that the change is in the direction of fixation. Considering multiple variants, a similar principle applies where the more variants we have, and the closer to 50% they are, the more resolution we have to distinguish different models of transmission. This suggests a general principle wherein the larger the donor diversity, the more robust the transmission inference. However, where a large number of variant sites exists we reach a point where an exhaustive approach to haplotype reconstruction results in a large number of potential haplotypes, slowing down the inference substantially. Instead, this regime favours a more minimal approach to haplotype reconstruction such as the MLHapRec method of Chapter 2.

In some situations, neutral markers or molecular barcodes may be added to a viral population (Abel et al. 2015; Varble et al. 2014); without providing an estimate of selection, sequencing these markers before and after transmission can give a precise estimate of the population bottleneck. While our method does not require the presence of such markers, its adaptation to include marker data would likely be straightforward, including in a calculation a further probabilistic term constraining the bottleneck size. Inference of selection for transmissibility could then be conducted under this constraint; the combination of whole-genome sequence data with such information could prove powerful for the study of viral transmission.

While I have here considered the transmission of influenza virus, very few steps of the approach would need to be altered for the method to be applied to another viral population. As detailed in the Methods section, it is only in accounting for genetic drift in the within-host growth of the virus that I make approximations relying on biological knowledge of the influenza virus; an alternative accounting for within-host expansion could be used. A second key assumption in the inference of selection is the existence of regions of the virus separated from each other by recombination or reassortment. This assumption would be preserved in some other viruses, as noted in observations of within-host HIV evolution (Zanini et al. 2015), if not for all influenza populations (Sobel Leonard et al. 2017a). Where a viral genome did not exhibit recombination, and only a single transmission event was observed, the neutral version of the method could be applied; in this context our accounting for haplotype structure

and sequencing noise in transmission represents an advance over methods which ignore these factors.

Chapter 4

Advanced Transmission Inference Scheme and Application to Experimental Data

4.1 Introduction

In chapter 3 I described the basic transmission model and verified its behaviour when applied to simulated data. Benchmarking was performed with respect to the Sobel Leonard et al. method where our approach was found to be additionally flexible with respect to capturing potentially different within-host replication processes. In this chapter I extend the basic transmission scheme to allow for more general considerations of within-host processes, including selection for increased within-host adaptation and multiple rounds of viral growth. The ability of the method to infer selection across multiple loci is also evaluated and relevant limitations discussed. On the basis of the advanced inference scheme I apply our model to an experimental influenza transmission dataset in ferrets, previously analysed by Moncla et al. (2016).

4.1.1 Author Contributions

The work presented in this chapter was previously published in Lumby, Nene and Illingworth (2018). The majority of the work described here was carried out by the author under the supervision of his PhD supervisor, Dr Christopher Illingworth. The inference of fitness landscapes was undertaken by Christopher Illingworth.

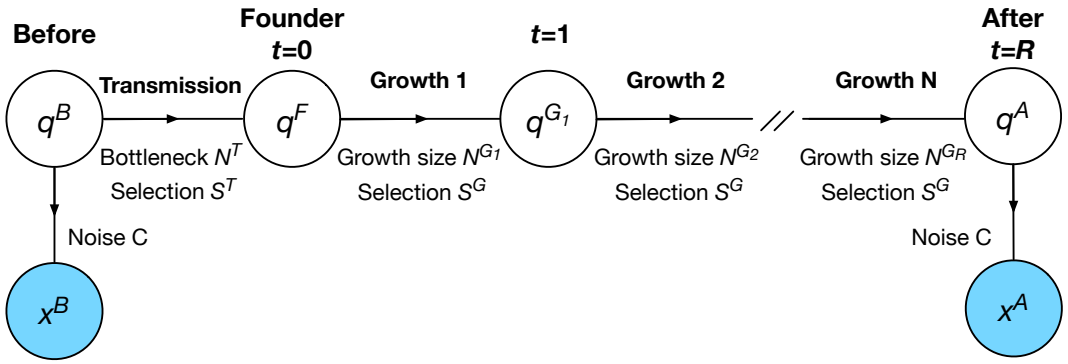


Figure 4.1. Probabilistic graph model for the advanced transmission inference scheme. This model extends the basic model given in Figure 3.2 by allowing for multiple rounds of within-host growth and incorporating selection for increased within-host adaptation, denoted by S^G.

4.2 Methods

4.2.1 Generalised Model of Transmission

I here outline our generalised model of transmission as seen in Figure 4.1. This model considers a general R-step within-host replication process accounting for drift and selection for increased within-host adaptation. Given a founder population size of N^T and a growth factor of g, the population size $N(t)$ assumes values $N(t) = N^T g^t$ after $t = 1, 2, \ldots, R$ replication cycles. Additionally we define $\boldsymbol{q}^A \equiv \boldsymbol{q}^{G_R}$, i.e. sampling takes place after R rounds of replication. The specific composition of the viral population at each time point is governed by a multinomial sampling process in the viral frequencies at the preceding time point. Selection acting for within-host growth may further alter the genetic composition of the population; this effect is described by the function S^G, acting once every replication cycle, and is independent of selection for increased transmissibility (Coombs, Gilchrist and Ball 2007). S^G is identical in form to S^T and responsible for changing the frequencies of haplotypes according to their relative propensity for within-host adaptation. Neglecting this effect could distort the inferred value of S^T; given only data collected before and after transmission the two terms cannot be separated. However, where samples have been collected at distinct times from one or multiple hosts, it is possible to make an independent estimate of S^G (Illingworth 2015), such that the two forms of selection can be discriminated.

4.2.2 Case of $R = 1$

Initially we consider the simpler case of $R = 1$, i.e. $\boldsymbol{q}^A \equiv \boldsymbol{q}^{G_R} = \boldsymbol{q}^{G_1}$. This scenario differs from the basic transmission model only in the final sampling step which we represent as a multivariate normal with mean (cf. Equation 3.21)

$$\mathrm{E}[\boldsymbol{x}_l^{A,\mathrm{P}}|\boldsymbol{q}^A] = N_l^A T_l S^G(\boldsymbol{q}^A) \tag{4.1}$$

and variance (cf. Equation 3.22)

$$\mathrm{var}[\boldsymbol{x}_l^{A,\mathrm{P}}|\boldsymbol{q}^A] = \left(\frac{N_l^A + C}{1 + C}\right) N_l^A M(T_l S^G(\boldsymbol{q}^A)) \equiv \alpha_l N_l^A M(T_l S^G(\boldsymbol{q}^A)) \tag{4.2}$$

where $\alpha_l = \left(\frac{N_l^A + C}{1+C}\right)$ and the subscript l denotes a specific partial haplotype set. The within-host growth selection function is defined in a manner identical to that of S^T (Equation 3.15):

$$\left(S^G(\boldsymbol{q}^A)\right)_i = \frac{w_i^G q_i^A}{\sum_{i'} w_{i'}^G q_{i'}^A} \tag{4.3}$$

where $\boldsymbol{w}^G = \{w_i^G\}$ are this within-host haplotype fitnesses. Under the assumption of two rounds of replication per 24 hours (see Section 4.3.1.2), inferred selection coefficients (in 12-hour units) were doubled in the computation of w_i^G (i.e. $s_k \to 2s_k$ in Equation 3.13). Selection is here assumed to act on the population after the increase in population size.

Based on the above we may rederive the resulting distribution for the $\boldsymbol{x}^A$ component, accounting for selection for increased within-host adaptation. Recalling the mean and variance for the after population (Equation 3.27 and 3.28 respectively):

$$\mathrm{E}[\boldsymbol{q}^A] = S^T(\boldsymbol{\mu}^B) \tag{4.4}$$

and

$$\mathrm{var}(\boldsymbol{q}^A) = \gamma M\left(S^T(\boldsymbol{\mu}^B)\right) + \delta\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)\Sigma^B\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger \tag{4.5}$$

where $\gamma = \left(\frac{N^T + N^G - 1}{N^T N^G}\right)$ and $\delta = \frac{N^T N^G - N^T - N^G + 1}{N^T N^G}$.

Performing the compound over $\boldsymbol{q}^A$ we obtain by the law of total expectation:

$$\begin{aligned}\mathrm{E}[\boldsymbol{x}_l^{A,\mathrm{P}}] &= \mathrm{E}[\mathrm{E}[\boldsymbol{x}_l^{A,\mathrm{P}}|\boldsymbol{q}^A]] = \mathrm{E}[N_l^A T_l S^G(\boldsymbol{q}^A)] \\ &\approx N_l^A T_l S^G\left(\mathrm{E}[\boldsymbol{q}^A]\right) = N_l^A T_l S^G\left(S^T\left(\boldsymbol{\mu}^B\right)\right)\end{aligned} \tag{4.6}$$

where we used the first-order second-moment method. Analogously, the law of total variance yields

$$\begin{aligned}\mathrm{var}(\boldsymbol{x}_l^{A,\mathrm{P}}) &= \mathrm{E}[\mathrm{var}[\boldsymbol{x}_l^{A,\mathrm{P}}|\boldsymbol{q}^A]] + \mathrm{var}[\mathrm{E}[\boldsymbol{x}_l^{A,\mathrm{P}}|\boldsymbol{q}^A]] \\ &= \mathrm{E}\left[\alpha_l N_l^A M\left(T_l S^G\left(\boldsymbol{q}^A\right)\right)\right] + \mathrm{var}[N_l^A T_l S^G\left(\boldsymbol{q}^A\right)] \\ &= \alpha_l N_l^A \left(\mathrm{Diag}(\mathrm{E}\left[T_l S^G\left(\boldsymbol{q}^A\right)\right] - \mathrm{E}\left[T_l S^G\left(\boldsymbol{q}^A\right)\right]\mathrm{E}\left[T_l S^G\left(\boldsymbol{q}^A\right)\right]^\dagger\right) \\ &\quad + N_l^A\left(N_l^A - \alpha_l\right)\mathrm{var}[T_l S^G\left(\boldsymbol{q}^A\right)] \\ &\approx \alpha_l N_l^A \left(\mathrm{Diag}(T_l S^G\left(\mathrm{E}\left[\boldsymbol{q}^A\right]\right) - T_l S^G\left(\mathrm{E}\left[\boldsymbol{q}^A\right]\right)\left(T_l S^G\left(\mathrm{E}\left[\boldsymbol{q}^A\right]\right)\right)^\dagger\right) \\ &\quad + N_l^A\left(N_l^A - \alpha_l\right) T_l \left(DS^G\big|_{\mathrm{E}[\boldsymbol{q}^A]}\right)\mathrm{var}\left[\boldsymbol{q}^A\right]\left(DS^G\big|_{\mathrm{E}[\boldsymbol{q}^A]}\right)^\dagger T_l^\dagger \\ &= \alpha_l N_l^A M(T_l S^G(S^T(\boldsymbol{\mu}^B))) + N_l^A\left(N_l^A - \alpha_l\right) T_l \left(DS^G\big|_{S^T(\boldsymbol{\mu}^B)}\right) \times \\ &\quad \left(\gamma M\left(S^T(\boldsymbol{\mu}^B)\right) + \delta\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)\Sigma^B\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger\right)\left(DS^G\big|_{S^T(\boldsymbol{\mu}^B)}\right)^\dagger T_l^\dagger\end{aligned} \tag{4.7}$$

The above mean and variance define the resulting distribution for the post-transmission sampling event in the presence of within-host selection assuming a single round of within-host growth. Notice that these expressions collapse to the special case of Equations 3.29 and 3.30 in the absence of within-host selection (see also Section 3.2.4.5).

4.2.3 Case of $R > 1$

In the case of $R > 1$ we may consider four different scenarios: A) a neutral transmission process, B) a transmission process under selection for transmission, C) a transmission process under selection for within-host evolution, and D) a transmission process under selection for both increased transmissibility and within-host adaptation. I here derive compound distributions for scenarios A) and B), which have closed form solutions, and refer to Appendix D for derivations for scenarios C) and D), which have recursive solutions.

4.2.3.1 Scenario A: Neutral Transmission Event

Considering a neutral transmission process we may define conditional expectations and variances for the random variables depicted in Figure 4.1. The pre-transmission process, defining the likelihood for $\boldsymbol{\mu}^B$ and Σ^B (Equation 3.5), is unchanged. The founder population is defined by

$$\mathrm{E}[\boldsymbol{q}^F|\boldsymbol{q}^B] = \boldsymbol{q}^B \tag{4.8}$$

and

$$\mathrm{var}[\boldsymbol{q}^F|\boldsymbol{q}^B] = \frac{1}{N^T} M(\boldsymbol{q}^B) \tag{4.9}$$

The initial growth step has conditional mean and variance

$$\mathrm{E}[\boldsymbol{q}^{G_1}|\boldsymbol{q}^F] = \boldsymbol{q}^F \tag{4.10}$$

and

$$\mathrm{var}[\boldsymbol{q}^{G_1}|\boldsymbol{q}^F] = \frac{1}{N^{G_1}} M(\boldsymbol{q}^F) = \frac{1}{gN^T} M(\boldsymbol{q}^F) \tag{4.11}$$

where $N^{G_1} = gN^T$.

In general, the nth growth step is defined by

$$\mathrm{E}[\boldsymbol{q}^{G_n}|\boldsymbol{q}^{G_{n-1}}] = \boldsymbol{q}^{G_{n-1}} \tag{4.12}$$

and

$$\mathrm{var}[\boldsymbol{q}^{G_n}|\boldsymbol{q}^{G_{n-1}}] = \frac{1}{N^{G_n}} M(\boldsymbol{q}^{G_{n-1}}) = \frac{1}{g^n N^T} M(\boldsymbol{q}^{G_{n-1}}) \tag{4.13}$$

where $N^{G_n} = g^n N^T$ and $n > 1$.

Assuming R steps in the growth process (i.e. $\boldsymbol{q}^A = \boldsymbol{q}^{G_R}$), we have

$$\mathrm{E}[\boldsymbol{q}^A|\boldsymbol{q}^{G_{R-1}}] = \boldsymbol{q}^{G_{R-1}} \tag{4.14}$$

and

$$\mathrm{var}[\boldsymbol{q}^A|\boldsymbol{q}^{G_{R-1}}] = \frac{1}{g^R N^T} M(\boldsymbol{q}^{G_{R-1}}) \tag{4.15}$$

As previously, the post-transmission observation results from a Dirichlet-multinomial sampling event:

$$\mathrm{E}[\boldsymbol{x}_i^{A,\mathrm{P}}|\boldsymbol{q}^A] = N_i^A T_i \boldsymbol{q}^A \tag{4.16}$$

and

$$\mathrm{var}[\boldsymbol{x}_i^{A,\mathrm{P}}|\boldsymbol{q}^A] = \alpha_i N_i^A M(T_i \boldsymbol{q}^A) \tag{4.17}$$

where $\alpha_i = \frac{N_i^A + C}{1+C}$.

Given the above conditional distributions we may perform the relevant marginalisations. The integral over $\boldsymbol{q}^B$ yields

$$\mathrm{E}[\boldsymbol{q}^F] = \mathrm{E}[\mathrm{E}[\boldsymbol{q}^F|\boldsymbol{q}^B]] = \mathrm{E}[\boldsymbol{q}^B] = \boldsymbol{\mu}^B \tag{4.18}$$

and,

$$\begin{aligned}
\mathrm{var}(\boldsymbol{q}^F) &= \mathrm{E}[\mathrm{var}[\boldsymbol{q}^F|\boldsymbol{q}^B]] + \mathrm{var}[\mathrm{E}[\boldsymbol{q}^F|\boldsymbol{q}^B]] \\
&= \mathrm{E}\left[\frac{1}{N^T} M(\boldsymbol{q}^B)\right] + \mathrm{var}\left[\boldsymbol{q}^B\right] \\
&= \frac{1}{N^T} M(\mathrm{E}[\boldsymbol{q}^B]) + \left(1 - \frac{1}{N^T}\right) \mathrm{var}[\boldsymbol{q}^B] \\
&= \frac{1}{N^T} M(\boldsymbol{\mu}^B) + \left(1 - \frac{1}{N^T}\right) \Sigma^B \\
&= \gamma_0 M(\boldsymbol{\mu}^B) + \delta_0 \Sigma^B
\end{aligned} \tag{4.19}$$

where in the last step we defined $\gamma_0 = \frac{1}{N^T}$ and $\delta_0 = \left(1 - \frac{1}{N^T}\right)$.

Next, for the $\boldsymbol{q}^F$ integral, the law of total expectation yields

$$\mathrm{E}[\boldsymbol{q}^{G_1}] = \mathrm{E}[\mathrm{E}[\boldsymbol{q}^{G_1}|\boldsymbol{q}^F]] = \mathrm{E}[\boldsymbol{q}^F] = \boldsymbol{\mu}^B \tag{4.20}$$

Next, under the law of total variance,

$$\begin{aligned}
\mathrm{var}(\boldsymbol{q}^{G_1}) &= \mathrm{E}[\mathrm{var}[\boldsymbol{q}^{G_1}|\boldsymbol{q}^F]] + \mathrm{var}[\mathrm{E}[\boldsymbol{q}^{G_1}|\boldsymbol{q}^F]] \\
&= \mathrm{E}\left[\frac{1}{gN^T}\left(\mathrm{Diag}(\boldsymbol{q}^F) - \boldsymbol{q}^F(\boldsymbol{q}^F)^\dagger\right)\right] + \mathrm{var}[\boldsymbol{q}^F] \\
&= \frac{1}{gN^T}\left(\mathrm{Diag}(\mathrm{E}[\boldsymbol{q}^F]) - \mathrm{E}[\boldsymbol{q}^F]E[\boldsymbol{q}^F]^\dagger\right) + \left(1 - \frac{1}{gN^T}\right) \mathrm{var}[\boldsymbol{q}^F] \\
&= \frac{1}{gN^T} M\left(\boldsymbol{\mu}^B\right) + \left(1 - \frac{1}{gN^T}\right)\left(\gamma_0 M(\boldsymbol{\mu}^B) + \delta_0 \Sigma^B\right) \\
&= \left(\frac{1}{gN^T} + \left(1 - \frac{1}{gN^T}\right)\gamma_0\right) M\left(\boldsymbol{\mu}^B\right) + \left(1 - \frac{1}{gN^T}\right)\delta_0 \Sigma^B \\
&\equiv \gamma_1 M\left(\boldsymbol{\mu}^B\right) + \delta_1 \Sigma^B
\end{aligned} \tag{4.21}$$

where we defined $\gamma_1 = \frac{1}{gN^T} + \left(1 - \frac{1}{gN^T}\right)\gamma_0$ and $\delta_1 = \left(1 - \frac{1}{gN^T}\right)\delta_0$.

Continuing with the marginalisation over $\boldsymbol{q}^{\boldsymbol{G_1}}$:

$$\mathrm{E}[\boldsymbol{q}^{G_2}] = \mathrm{E}[\mathrm{E}[\boldsymbol{q}^{G_2}|\boldsymbol{q}^{G_1}]] = \mathrm{E}[\boldsymbol{q}^{G_1}] = \boldsymbol{\mu}^B \tag{4.22}$$

and

$$\begin{aligned}
\mathrm{var}(\boldsymbol{q}^{G_2}) &= \mathrm{E}[\mathrm{var}[\boldsymbol{q}^{G_2}|\boldsymbol{q}^{G_1}]] + \mathrm{var}[\mathrm{E}[\boldsymbol{q}^{G_2}|\boldsymbol{q}^{G_1}]] \\
&= \mathrm{E}\left[\frac{1}{g^2N^T}\left(\mathrm{Diag}(\boldsymbol{q}^{G_1}) - \boldsymbol{q}^{G_1}(\boldsymbol{q}^{G_1})^\dagger\right)\right] + \mathrm{var}[\boldsymbol{q}^{G_1}] \\
&= \frac{1}{g^2N^T}\left(\mathrm{Diag}(\mathrm{E}[\boldsymbol{q}^{G_1}]) - \mathrm{E}[\boldsymbol{q}^{G_1}]E[\boldsymbol{q}^{G_1}]^\dagger\right) + \left(1 - \frac{1}{g^2N^T}\right)\mathrm{var}[\boldsymbol{q}^{G_1}] \\
&= \frac{1}{g^2N^T}M\left(\boldsymbol{\mu}^B\right) + \left(1 - \frac{1}{g^2N^T}\right)\left(\gamma_1 M\left(\boldsymbol{\mu}^B\right) + \delta_1\Sigma^B\right) \\
&= \left(\frac{1}{g^2N^T} + \left(1 - \frac{1}{g^2N^T}\right)\gamma_1\right)M\left(\boldsymbol{\mu}^B\right) + \left(1 - \frac{1}{g^2N^T}\right)\delta_1\Sigma^B \\
&\equiv \gamma_2 M\left(\boldsymbol{\mu}^B\right) + \delta_2\Sigma^B
\end{aligned} \tag{4.23}$$

where in the last step we defined $\gamma_2 = \frac{1}{g^2N^T} + \left(1 - \frac{1}{g^2N^T}\right)\gamma_1$ and $\delta_1 = \left(1 - \frac{1}{g^2N^T}\right)\delta_1$.

From the above it is clear that general mean and variance expressions may be derived for arbitrary $\boldsymbol{q}^{G_n}$ taking the form:

$$\mathrm{E}[\boldsymbol{q}^{G_n}] = \boldsymbol{\mu}^B \tag{4.24}$$

$$\mathrm{var}(\boldsymbol{q}^{G_n}) = \gamma_n M\left(\boldsymbol{\mu}^B\right) + \delta_n\Sigma^B \tag{4.25}$$

where γ_n and δ_n obey the recurrence relations:

$$\gamma_n = \frac{1}{g^nN^T} + \left(1 - \frac{1}{g^nN^T}\right)\gamma_{n-1} \tag{4.26}$$

$$\delta_n = \left(1 - \frac{1}{g^nN^T}\right)\delta_{n-1} \tag{4.27}$$

with $\gamma_0 = \frac{1}{N^T}$ and $\delta_0 = \left(1 - \frac{1}{N^T}\right)$.

Analytical solutions may be found:

$$\gamma_n = \frac{(-1)^n g^{-\frac{n^2}{2}-\frac{n}{2}} \left(N^T\right)^{-n-1} (gN^T;g)_n \left(g\left(N^T\right)^2 \left(\sum_{j=0}^{n-1} \frac{(gN^T-1)(-1)^{1-j} g^{\frac{j^2}{2}+\frac{j}{2}-1} \left(N^T\right)^{j-1}}{(gN^T;g)_{j+1}} \right) + gN^T - 1 \right)}{gN^T - 1} \tag{4.28}$$

and

$$\delta_n = (-1)^n g^{-\frac{n^2}{2}-\frac{n}{2}} \left(N^T\right)^{-n-1} \left(N^T - 1\right) (gN^T;g)_n \tag{4.29}$$

where $(a;q)_n = \prod_{k=0}^{n-1}(1-aq^k)$ is the q-Pochhammer symbol. These expressions can be readily verified for low n.

To this end, assuming R steps in the growth process, the mean and variance of $\boldsymbol{q}^A$ are those of $\boldsymbol{q}^{G_R}$:

$$\mathrm{E}[\boldsymbol{q}^A] = \boldsymbol{\mu}^B \tag{4.30}$$

$$\mathrm{var}(\boldsymbol{q}^A) = \gamma_R M\left(\boldsymbol{\mu}^B\right) + \delta_R \Sigma^B \tag{4.31}$$

Finally we may compute the marginalisation over $\boldsymbol{q}^A$:

$$\mathrm{E}[\boldsymbol{x}_i^{A,\mathrm{P}}] = \mathrm{E}[\mathrm{E}[\boldsymbol{x}_i^{A,\mathrm{P}}|\boldsymbol{q}^A]] = \mathrm{E}[N_i^A T_i \boldsymbol{q}^A] = N_i^A T_i \mathrm{E}[\boldsymbol{q}^A] = N_i^A T_i \boldsymbol{\mu}^B \tag{4.32}$$

and

$$\begin{aligned}
\mathrm{var}(\boldsymbol{x}_i^{A,\mathrm{P}}) &= \mathrm{E}[\mathrm{var}[\boldsymbol{x}_i^{A,\mathrm{P}}|\boldsymbol{q}^A]] + \mathrm{var}[\mathrm{E}[\boldsymbol{x}_i^{A,\mathrm{P}}|\boldsymbol{q}^A]] \\
&= \mathrm{E}\left[\alpha N_i^A M\left(T_i \boldsymbol{q}^A\right)\right] + \mathrm{var}[N_i^A T_i \boldsymbol{q}^A] \\
&= \alpha_i N_i^A \Big(\mathrm{Diag}(\mathrm{E}\left[T_i \boldsymbol{q}^A\right] - \mathrm{E}\left[T_i \boldsymbol{q}^A\right] \mathrm{E}\left[T_i \boldsymbol{q}^A\right]^\dagger\Big) \\
&\qquad + N_i^A \left(N_i^A - \alpha_i\right) \mathrm{var}[T_i \boldsymbol{q}^A] \\
&= \alpha_i N_i^A \Big(\mathrm{Diag}\left(T_i \mathrm{E}\left[\boldsymbol{q}^A\right]\right) - T_i \mathrm{E}\left[\boldsymbol{q}^A\right] \left(T_i \mathrm{E}\left[\boldsymbol{q}^A\right]\right)^\dagger\Big) \\
&\qquad + N_i^A \left(N_i^A - \alpha_i\right) T_i \mathrm{var}\left[\boldsymbol{q}^A\right] T_i^\dagger \\
&= \alpha_i N_i^A M\left(T_i \boldsymbol{\mu}^B\right)) + N_i^A \left(N_i^A - \alpha_i\right) T_i \left(\gamma_R M\left(\boldsymbol{\mu}^B\right) + \delta_R \Sigma^B\right) T_i^\dagger \\
&= N_i^A \left(\alpha_i + \left(N_i^A - \alpha_i\right) \gamma_R\right) M\left(T_i \boldsymbol{\mu}^B\right) + N_i^A \left(N_i^A - \alpha_i\right) \delta_R T_i \Sigma^B T_i^\dagger
\end{aligned} \tag{4.33}$$

where in the last step we used that $T_i \mathrm{Diag}(\boldsymbol{\mu}^B) T_i^\dagger = \mathrm{Diag}(T_i \boldsymbol{\mu}^B)$ which is true if T_i consists of zeroes and ones and if every column of T_i contains a single non-zero element, i.e. if a full haplotype can only contribute to a single partial haplotype in the set i. See Appendix E for proof of this identity.

We note that the expressions for γ_n and δ_n may be generalised for arbitrary γ_0 and δ_0:

$$\gamma_n = \frac{(-1)^n g^{-\frac{n^2}{2}-\frac{n}{2}} \left(N^T\right)^{-n} (gN^T; g)_n \left(gN^T \left(\sum_{j=0}^{n-1} \frac{(gN^T-1)(-1)^{1-j} g^{\frac{j^2}{2}+\frac{j}{2}-1} \left(N^T\right)^{j-1}}{(gN^T;g)_{j+1}} \right) + \gamma_0 gN^T - \gamma_0 \right)}{gN^T - 1} \tag{4.34}$$

$$\delta_n = (-1)^n \delta_0 g^{-\frac{n^2}{2}-\frac{n}{2}} \left(N^T\right)^{-n} (gN^T; g)_n \tag{4.35}$$

4.2.3.2 Scenario B: Selection for Transmission

Compound distributions for transmission processes under selection for increased transmissibility may be derived in a manner similar to the above. The main difference arises from the conditional mean and variance of the founder population:

$$\mathrm{E}[\boldsymbol{q}^F | \boldsymbol{q}^B] = S^T(\boldsymbol{q}^B) \tag{4.36}$$

and

$$\mathrm{var}[\boldsymbol{q}^F | \boldsymbol{q}^B] = \frac{1}{N^T} M(S^T(\boldsymbol{q}^B)) \tag{4.37}$$

Starting with the marginalisation over $\boldsymbol{q}^B$ we obtain

$$\mathrm{E}[\boldsymbol{q}^F] = \mathrm{E}[\mathrm{E}[\boldsymbol{q}^F | \boldsymbol{q}^B]] = \mathrm{E}[S^T(\boldsymbol{q}^B)] \approx S^T(\mathrm{E}[\boldsymbol{q}^B]) = S^T(\boldsymbol{\mu}^B) \tag{4.38}$$

where in the penultimate step we used the first-order second-moment approximation to a vector function acting on a random variable. The law of total variance yields

$$
\begin{aligned}
\mathrm{var}(\boldsymbol{q}^F) &= \mathrm{E}[\mathrm{var}[\boldsymbol{q}^F|\boldsymbol{q}^B]] + \mathrm{var}[\mathrm{E}[\boldsymbol{q}^F|\boldsymbol{q}^B]] \\
&= \mathrm{E}\left[\frac{1}{N^T} M(S^T(\boldsymbol{q}^B))\right] + \mathrm{var}\left[S^T(\boldsymbol{q}^B)\right] \\
&= \frac{1}{N^T} M(\mathrm{E}[S^T(\boldsymbol{q}^B)]) + \left(1 - \frac{1}{N^T}\right) \mathrm{var}[S^T(\boldsymbol{q}^B)] \\
&\approx \frac{1}{N^T} M(S^T(\mathrm{E}[\boldsymbol{q}^B])) + \left(1 - \frac{1}{N^T}\right) \left(DS^T\big|_{\mathrm{E}[\boldsymbol{q}^B]}\right) \mathrm{var}[\boldsymbol{q}^B] \left(DS^T\big|_{\mathrm{E}[\boldsymbol{q}^B]}\right)^\dagger \\
&= \frac{1}{N^T} M(S^T(\boldsymbol{\mu}^B)) + \left(1 - \frac{1}{N^T}\right) \left(DS^T\big|_{\boldsymbol{\mu}^B}\right) \Sigma^B \left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger \\
&= \gamma_0 M(S^T(\boldsymbol{\mu}^B)) + \delta_0 \left(DS^T\big|_{\boldsymbol{\mu}^B}\right) \Sigma^B \left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger
\end{aligned}
\tag{4.39}
$$

where in the last step we defined $\gamma_0 = \frac{1}{N^T}$ and $\delta_0 = \left(1 - \frac{1}{N^T}\right)$. As previously, $(DS)_i^j = \frac{\partial S_i}{\partial q_j}$ is the Jacobian matrix arising from the first-order second-moment approximation.

Next, for the $\boldsymbol{q}^F$ integral, the law of total expectation gives

$$
\mathrm{E}[\boldsymbol{q}^{G_1}] = \mathrm{E}[\mathrm{E}[\boldsymbol{q}^{G_1}|\boldsymbol{q}^F]] = \mathrm{E}[\boldsymbol{q}^F] = S^T(\boldsymbol{\mu}^B) \tag{4.40}
$$

with the law of total variance yielding

$$
\begin{aligned}
\mathrm{var}(\boldsymbol{q}^{G_1}) &= \mathrm{E}[\mathrm{var}[\boldsymbol{q}^{G_1}|\boldsymbol{q}^F]] + \mathrm{var}[\mathrm{E}[\boldsymbol{q}^{G_1}|\boldsymbol{q}^F]] \\
&= \mathrm{E}\left[\frac{1}{gN^T}\left(\mathrm{Diag}(\boldsymbol{q}^F) - \boldsymbol{q}^F(\boldsymbol{q}^F)^\dagger\right)\right] + \mathrm{var}[\boldsymbol{q}^F] \\
&= \frac{1}{gN^T}\left(\mathrm{Diag}(\mathrm{E}[\boldsymbol{q}^F]) - \mathrm{E}[\boldsymbol{q}^F]E[\boldsymbol{q}^F]^\dagger\right) + \left(1 - \frac{1}{gN^T}\right) \mathrm{var}[\boldsymbol{q}^F] \\
&= \frac{1}{gN^T} M\left(S^T(\boldsymbol{\mu}^B)\right) + \left(1 - \frac{1}{gN^T}\right) \Big(\gamma_0 M(S^T(\boldsymbol{\mu}^B)) \\
&\qquad\qquad + \delta_0 \left(DS^T\big|_{\boldsymbol{\mu}^B}\right) \Sigma^B \left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger\Big) \\
&= \left(\frac{1}{gN^T} + \left(1 - \frac{1}{gN^T}\right)\gamma_0\right) M\left(S^T(\boldsymbol{\mu}^B)\right) \\
&\qquad\qquad + \left(1 - \frac{1}{gN^T}\right) \delta_0 \left(DS^T\big|_{\boldsymbol{\mu}^B}\right) \Sigma^B \left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger \\
&\equiv \gamma_1 M\left(S^T(\boldsymbol{\mu}^B)\right) + \delta_1 \left(DS^T\big|_{\boldsymbol{\mu}^B}\right) \Sigma^B \left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger
\end{aligned}
\tag{4.41}
$$

where we defined $\gamma_1 = \frac{1}{gN^T} + \left(1 - \frac{1}{gN^T}\right)\gamma_0$ and $\delta_1 = \left(1 - \frac{1}{gN^T}\right)\delta_0$.

Continuing with the marginalisation over $\boldsymbol{q}^{\boldsymbol{G_1}}$:

$$\mathrm{E}[\boldsymbol{q}^{G_2}] = \mathrm{E}[\mathrm{E}[\boldsymbol{q}^{G_2}|\boldsymbol{q}^{G_1}]] = \mathrm{E}[\boldsymbol{q}^{G_1}] = S^T(\boldsymbol{\mu}^B) \tag{4.42}$$

and

$$\begin{aligned}
\mathrm{var}(\boldsymbol{q}^{G_2}) &= \mathrm{E}[\mathrm{var}[\boldsymbol{q}^{G_2}|\boldsymbol{q}^{G_1}]] + \mathrm{var}[\mathrm{E}[\boldsymbol{q}^{G_2}|\boldsymbol{q}^{G_1}]] \\
&= \mathrm{E}\left[\frac{1}{g^2N^T}\left(\mathrm{Diag}(\boldsymbol{q}^{G_1}) - \boldsymbol{q}^{G_1}(\boldsymbol{q}^{G_1})^\dagger\right)\right] + \mathrm{var}[\boldsymbol{q}^{G_1}] \\
&= \frac{1}{g^2N^T}\left(\mathrm{Diag}(\mathrm{E}[\boldsymbol{q}^{G_1}]) - \mathrm{E}[\boldsymbol{q}^{G_1}]E[\boldsymbol{q}^{G_1}]^\dagger\right) + \left(1 - \frac{1}{g^2N^T}\right)\mathrm{var}[\boldsymbol{q}^{G_1}] \\
&= \frac{1}{g^2N^T}M\left(S^T(\boldsymbol{\mu}^B)\right) + \left(1 - \frac{1}{g^2N^T}\right)\left(\gamma_1 M\left(S^T(\boldsymbol{\mu}^B)\right)\right. \\
&\qquad \left. + \delta_1\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)\Sigma^B\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger\right) \\
&= \left(\frac{1}{g^2N^T} + \left(1 - \frac{1}{g^2N^T}\right)\gamma_1\right)M\left(S^T(\boldsymbol{\mu}^B)\right) \\
&\qquad + \left(1 - \frac{1}{g^2N^T}\right)\delta_1\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)\Sigma^B\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger \\
&\equiv \gamma_2 M\left(S^T(\boldsymbol{\mu}^B)\right) + \delta_2\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)\Sigma^B\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger
\end{aligned} \tag{4.43}$$

where in the last step we defined $\gamma_2 = \frac{1}{g^2N^T} + \left(1 - \frac{1}{g^2N^T}\right)\gamma_1$ and $\delta_2 = \left(1 - \frac{1}{g^2N^T}\right)\delta_1$.

From the above it is clear that general mean and variance expressions may defined for arbitrary $\boldsymbol{q}^{G_n}$:

$$\mathrm{E}[\boldsymbol{q}^{G_n}] = S^T(\boldsymbol{\mu}^B) \tag{4.44}$$

$$\mathrm{var}(\boldsymbol{q}^{G_n}) = \gamma_n M\left(S^T(\boldsymbol{\mu}^B)\right) + \delta_n\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)\Sigma^B\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger \tag{4.45}$$

where γ_n and δ_n obey the recurrence relations:

$$\gamma_n = \frac{1}{g^nN^T} + \left(1 - \frac{1}{g^nN^T}\right)\gamma_{n-1} \tag{4.46}$$

$$\delta_n = \left(1 - \frac{1}{g^n N^T}\right) \delta_{n-1} \tag{4.47}$$

with $\gamma_0 = \frac{1}{N^T}$ and $\delta_0 = \left(1 - \frac{1}{N^T}\right)$. The solutions to these recurrence relations are given in Equations 4.28 and 4.29.

To this end, assuming R steps in the growth process, the mean and variance of $\boldsymbol{q}^A$ are those of $\boldsymbol{q}^{G_R}$:

$$\mathrm{E}[\boldsymbol{q}^A] = S^T(\boldsymbol{\mu}^B) \tag{4.48}$$

$$\mathrm{var}(\boldsymbol{q}^A) = \gamma_R M\left(S^T(\boldsymbol{\mu}^B)\right) + \delta_R \left(DS^T\big|_{\boldsymbol{\mu}^B}\right) \Sigma^B \left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger \tag{4.49}$$

Finally we may compute the marginalisation over $\boldsymbol{q}^A$:

$$\mathrm{E}[\boldsymbol{x}_i^{A,\mathrm{P}}] = \mathrm{E}[\mathrm{E}[\boldsymbol{x}_i^{A,\mathrm{P}}|\boldsymbol{q}^A]] = \mathrm{E}[N_i^A T_i \boldsymbol{q}^A] = N_i^A T_i \mathrm{E}[\boldsymbol{q}^A] = N_i^A T_i S^T(\boldsymbol{\mu}^B) \tag{4.50}$$

and

$$\begin{aligned}
\mathrm{var}(\boldsymbol{x}_i^{A,\mathrm{P}}) &= \mathrm{E}[\mathrm{var}[\boldsymbol{x}_i^{A,\mathrm{P}}|\boldsymbol{q}^A]] + \mathrm{var}[\mathrm{E}[\boldsymbol{x}_i^{A,\mathrm{P}}|\boldsymbol{q}^A]] \\
&= \mathrm{E}\left[\alpha N_i^A M\left(T_i \boldsymbol{q}^A\right)\right] + \mathrm{var}[N_i^A T_i \boldsymbol{q}^A] \\
&= \alpha_i N_i^A \Big(\mathrm{Diag}(\mathrm{E}\left[T_i \boldsymbol{q}^A\right] - \mathrm{E}\left[T_i \boldsymbol{q}^A\right] \mathrm{E}\left[T_i \boldsymbol{q}^A\right]^\dagger\Big) \\
&\qquad + N_i^A \left(N_i^A - \alpha_i\right) \mathrm{var}[T_i \boldsymbol{q}^A] \\
&= \alpha_i N_i^A \Big(\mathrm{Diag}\left(T_i \mathrm{E}\left[\boldsymbol{q}^A\right]\right) - T_i \mathrm{E}\left[\boldsymbol{q}^A\right] \left(T_i \mathrm{E}\left[\boldsymbol{q}^A\right]\right)^\dagger\Big) \\
&\qquad + N_i^A \left(N_i^A - \alpha_i\right) T_i \mathrm{var}\left[\boldsymbol{q}^A\right] T_i^\dagger \\
&= \alpha_i N_i^A M\left(T_i S^T(\boldsymbol{\mu}^B)\right)\big) + N_i^A \left(N_i^A - \alpha_i\right) T_i \Big(\gamma_R M\left(S^T(\boldsymbol{\mu}^B)\right) \\
&\qquad + \delta_R \left(DS^T\big|_{\boldsymbol{\mu}^B}\right) \Sigma^B \left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger\Big) T_i^\dagger \\
&= N_i^A \left(\alpha_i + \left(N_i^A - \alpha_i\right) \gamma_R\right) M\left(T_i S^T(\boldsymbol{\mu}^B)\right) \\
&\qquad + N_i^A \left(N_i^A - \alpha_i\right) \delta_R T_i \left(DS^T\big|_{\boldsymbol{\mu}^B}\right) \Sigma^B \left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger T_i^\dagger
\end{aligned} \tag{4.51}$$

where in the last step we used the identity $T_i \text{Diag}(S^T(\boldsymbol{\mu}^B))T_i^\dagger = \text{Diag}(T_i S^T(\boldsymbol{\mu}^B))$ (see Appendix E).

4.2.4 Analysis of Simulated Data

A number of more elaborate transmission processes were simulated and investigated on the basis of the advanced inference model.

4.2.4.1 Selection Inference in the Presence of Within-Host Selection

Inferring selection for transmission in the presence of selection for within-host growth is a difficult task. Differentiating between the two effects is impossible given transmission data from only two time points, however, with multiple available samples one may infer selection for within-host adaptation separately from that of selection for transmission. To determine the ability of our method to account for within-host selection, and to highlight the bias arising from the neglect of it, I simulated and analysed data from transmission events affected by within-host adaptive processes. I considered four different scenarios: A) the presence of selection for transmission only, B) the presence of selection for within-host growth only combined with an inference model not accounting for this, C) the presence of selection for within-host growth combined with an inference model which accounts for this, and D) the presence of selection for transmission and within-host growth using an inference model accounting for the within-host selection. Where present, selection for transmission was chosen to act on the third (of five) loci in HA with selection for within-host growth acting on the third loci in NA. Substantial selection of $\sigma^T = \sigma^{\text{WH}} = 1$ were applied in order to amplify the signal and highlight the general trends obtained.

4.2.4.2 Bottleneck Inference Under Multiple Rounds of Within-Host Growth

Following up on the results of Section 3.3.8 regarding the ability of the assumed within-host biology to impact inference outcomes, I sought to consider a more complicated, and perhaps more realistic, within-host growth scheme. I here simulated neutral transmission data incorporating a four-step within-host growth process, considering a range of growth factors, $g = \{1, 2, 5, 10, 22\}$, and the presence ($C = 200$) or absence ($C = 10^6$) of noise. Bottleneck inferences were

conducted using the basic model (accounting for a single round of within-host growth) and the advanced model (accounting for four rounds of growth).

4.2.4.3 Selection on Multiple Variants

In the previous chapter selection inference was considered only on the basis of a single variant conferring increased transmissibility. In order to determine the ability of our method to infer subsequent instances of selection, I evaluated simulated data with selection acting on two variant sites. I considered the difference in inferences when the selective variants were either located within a single gene segment or upon two different segments. Selection inferences were performed using a BIC penalty function derived using the process outlined in Section 3.2.15. Simulated data were generated as described in 3.2.12 with selection acting either on the second and fourth variant in HA (hereafter referred to as scenario A) or on the third variant site in HA and NA (scenario B).

4.2.5 Experimental Sequence Data

Data were analysed from an evolutionary experiment considering airborne transmission of a 1918-like influenza virus between ferrets (Moncla et al. 2016). The specific data examined here describe two sets of viral transmissions. In the first, denoted HA190D220D, a viral population was given to three ferrets, transmission to a recipient host being observed in one of three cases, giving time-resolved sequence data from four ferrets. In the second, denoted Mut, a viral population arising from the first experiment was given to three ferrets, transmission to two recipient hosts being observed, giving data from five ferrets.

4.2.6 Processing of Sequence Data

Genome sequence data was processed using the SAMFIRE software package, according to default settings (Illingworth 2016), calling variant alleles that existed at a frequency of at least 1% at some point during the observed infections. For the calculation of a within-host fitness landscape, the effective depth of sequencing was estimated in a conservative manner, filtering out variants which changed in frequency by more than 5% per day before using the frequencies of remaining variants from different time-points within the same host to estimate the parameter C. For the within-host model, following the approach of previous calculations (Illingworth 2015; Sobel Leonard et al. 2017a), potentially non-neutral

variants were identified as those for which a model of frequency change under selection outperformed a neutral model by more than 10 units according to the Bayesian Information Criterion (BIC) (Kass and Raftery 1995). Variants reaching a frequency of at least 5% in at least one sample were then identified before calling multi-locus variant observations from the data; data from all time-points for which within-host data were collected were used in this inference. The 5% cutoff was chosen to reduce computational costs for this part of the calculation while still reconstructing the core aspects of the within-host fitness landscape.

For the inference of transmission, data from all polymorphic sites were utilised, with no filtering of sites. As in the original analysis of the data (Moncla et al. 2016), variants were identified from data collected from the final observation before transmission and the first point of observation after transmission; these data were used to construct multi-locus observations across variants which reached a frequency of at least 2% in at least one sample. In this inference a revised approach to estimating the effective depth of sequencing was taken, noting our result that estimates which overestimate noise may lead to errors in the inferred bottleneck size. Here, in common with previous calculations, we initially identified a conservative value of C from within-host data using the default settings in SAMFIRE. Next, variant frequencies were evaluated, identifying potentially non-neutral changes in frequency using a single-locus analysis (Illingworth 2015). Finally, a more conservative estimate of C was calculated, using the set of trajectories which were identified as being consistent with a neutral model of frequency change. This conservative estimate reflects our finding (shown in Chapter 3) that a conservative estimate of noise (i.e. potentially underestimating the noise) is less likely to induce a substantial error in the inferred bottleneck than a less conservative approach, which might overestimate the extent of noise in the data.

Additional processing of transmission data was carried out within the transmission inference code according to the framework outlined in Section 3.2.13.

4.2.7 Inference of Within-Host Selection

For the experimental dataset an inference of within-host selection was conducted according to a method previously described in earlier publications (Illingworth 2015; Sobel Leonard et al. 2017a). Under the assumption of rapid reassortment in the system (Marshall et al. 2013) different segments of the virus were treated independently. Our inference of selection aimed to characterise fitness so as

to estimate S^G for an inference of transmission; the HA190D225D and Mut datasets were considered independently, with data from all animals in each set being combined to infer within-host selection.

Our approach to the inference of selection is similar to that for transmission, yet works under the assumption of a large population size, and may involve data from more than two time points. Taking data from a set of animals a hierarchical approach is again taken to the inference of selection, testing a neutral model of evolution against successively more complex within-host fitness landscapes, such landscapes being defined as the sum of single-locus and multi-locus (epistatic) fitness effects. In this calculation, the application of BIC for model selection is more straightforward; as more parameters are added into the model the fitness landscape becomes incrementally more complex, but in a convergent manner: Statistically each additional parameter makes a smaller change to the landscape as a whole (Illingworth 2015). As such, a fixed threshold of 10 BIC units was required to accept the addition of a further parameter.

4.3 Results

4.3.1 Application to Simulated Data

I investigated the performance of the advanced transmission model on simulated data, aiming on the one hand to consider inference of selection for transmission in the presence of within-host selection, and on the other to validate the model under more complicated within-host growth configurations. Scenarios involving selection for multiple polymorphic sites were also explored.

4.3.1.1 Selection for Within-Host Adaptation Bias the Inference of Selection for Transmission

Figure 4.2, and 4.3 consider the impact of selection for within-host adaptation on the inference of selection for transmission. Figure 4.2 shows the true and false positive rates of inference of selection for transmission for four distinct cases. The top left subplot corresponds to simulations for which only selection for increased transmissibility is present. This plot is identical to the case of $\sigma^T = 1$ in Figure 3.18. Next, the top right panel considers the presence of within-host selection in the absence of selection for transmission. The inference model aims to infer selection for transmission without correctly accounting for

the within-host selection ($\sigma^{\mathrm{WH,ana}} = 0$). The bottom left subplot corresponds to an identical simulation setup, but with the inference model accurately accounting for the presence of within-host selection ($\sigma^{\mathrm{WH,ana}} = 1$). Finally, the last subplot considers the presence of both selection for increased transmissibility and for within-host growth with the latter appropriately accounted for within the framework.

In the top right subplot we observe a substantial false positive rate of inference of selection for transmission compared to the baseline false positive rate (top left plot) which is less than 10%. This is due to the method attributing diversity changes arising from within-host selection to the action of selection during transmission. However, the false positive rate is lower than the baseline true positive rate, where, given that the magnitudes of selection are equal, we might expect these to be comparable. This effect can be explained by the difference in timings of the two selection impacts. Selection for transmission acts prior to the bottleneck event whilst selection for within-host growth acts subsequently to this. The population bottleneck reduces the overall diversity of the viral population, reducing the potential for further changes in diversity to provide the statistical support required to make an inference of selection.

In the bottom left plot we correctly account for within-host selection within the inference model; this results in a low false positive rate. This roughly corresponds to what would be expected from an inference of transmission in an event where no selection occurs. In the final scenario (bottom right) we observe the ability to infer true positive instances of selection for transmission even in the presence of within-host selection. We note that the true positive rate is marginally lower than the baseline rate; again we believe this can be understood in terms of the reduction in diversity from within-host selection limiting the potential to identify selection from other changes in this population.

In Figure 4.3 we investigate the inferred strengths of selection for transmission. The left plot considers the baseline case and is identical to the $\sigma^T = 1$ outcome in Figure 3.19. As noted previously, the probability density functions are centered around $\sigma^T = 1$ (or slightly above 1 as the bottleneck is decreased). In the presence of within-host selection we observe an almost identical set of distributions, highlighting the fact that the method not only infers the presence of selection for increased transmissibility but also captures the magnitude correctly. The areas under the curves are marginally smaller, reflecting that a lower fraction of true positives were inferred here compared to the baseline case. Taken together, Figure 4.2, and 4.3 demonstrate the importance of accounting

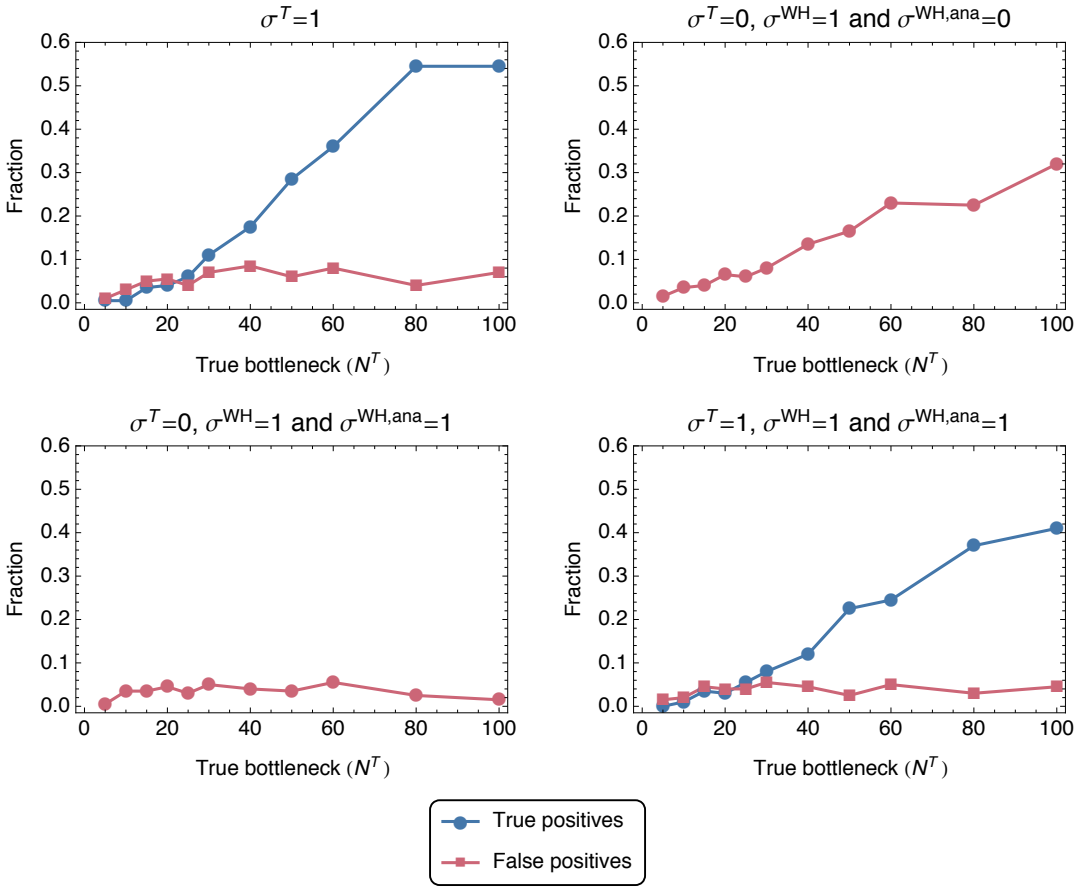


Figure 4.2. True and false positive rates of selection inference from 200 simulations of transmission events with differing types of selection present and accounted for. The four scenarios are as follows: presence of selection for transmission of strength $\sigma^T = 1$ (top left), presence of selection for within-host selection (not accounted for in model) of strength $\sigma^{\mathrm{WH}} = 1$ (top right), presence of selection for within-host selection (accounted for) of strength $\sigma^{\mathrm{WH}} = 1$ (bottom left), and presence of selection for transmission and within-host selection (accounted for) of strengths $\sigma^T = \sigma^{\mathrm{WH}} = 1$ (bottom right). True positives were defined as inferences for which selection for transmission was inferred for the selected locus; false positives were defined as inferences for which selection was inferred at any neutral locus or for multiple neutral loci in the system. Inferences can be simultaneously true and false positive (see e.g. Figure 3.10). Where present, selection for transmission was chosen to act on the third of five loci in HA and selection for within-host growth acting on the third of five loci in NA.

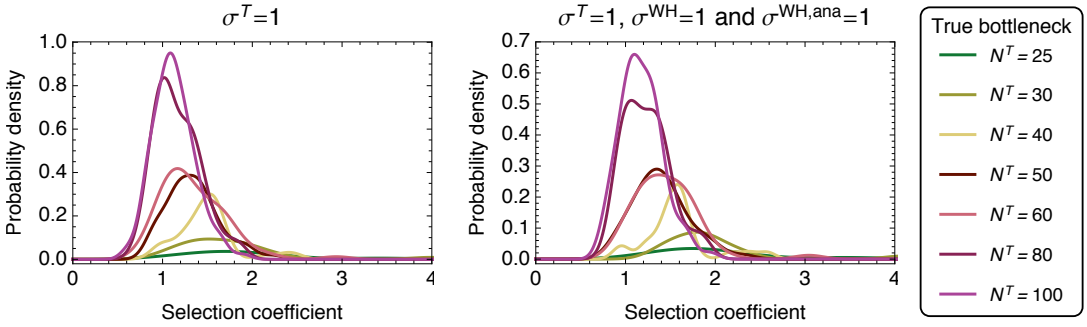


Figure 4.3. Probability distributions of inferred selection coefficients from 200 simulations of transmission events with selective pressures $\sigma^T = 1$ (left) and $\sigma^T = \sigma^{\mathrm{WH}} = 1$ (the latter accounted for in model, right). Distributions were constructed for bottleneck values where the inference of selection resulted in a true positive rate for identifying selected variants of above 5 %. Smooth kernel distributions were computed using a Gaussian kernel function defined on $(0, 10)$ and Silverman's rule of thumb (Silverman 1986, p. 48) employed for the bandwidth size. Distributions were scaled such that their integral across the kernel range equalled the true positive rate.

for different sources of selection and the ability of our framework to correctly capture this aspect when an estimate of within-host selection is available.

4.3.1.2 Advanced Model of Within-Host Evolution

In all previous simulation experiments I have considered a simple generative model for which within-host growth consists of a single round of replication. In experimental datasets the true within-host biology may differ from this; different viruses may have different number of within-host replication rounds or the recipient host may be sampled more sparsely leading to additional within-host evolution. In general the within-host growth process was simulated as repeated multinomial sampling for a total of

$$n_{\mathrm{rounds}} = n_{\mathrm{generations}} \cdot \Delta_{\mathrm{sampling}} \tag{4.52}$$

times, where $n_{\mathrm{generations}}$ is the number of cellular cycles per 24 hours and $\Delta_{\mathrm{sampling}}$ is the number of days between sampling times. Influenza virus is believed to have approximately two rounds of replication per day (Baccam et al. 2006; Russell et al. 2012; Sidorenko and Reichl 2004).

In Figure 4.4 I investigated the ability of the inference model to infer bottleneck sizes from neutral transmission events with within-host growth processes consisting of $n_{\mathrm{rounds}} = 4$ rounds of cellular cycles. I considered a range of growth factors ($g = \{1, 2, 5, 10, 22\}$) and noise values ($C = \{200, 10^6\}$) and compared

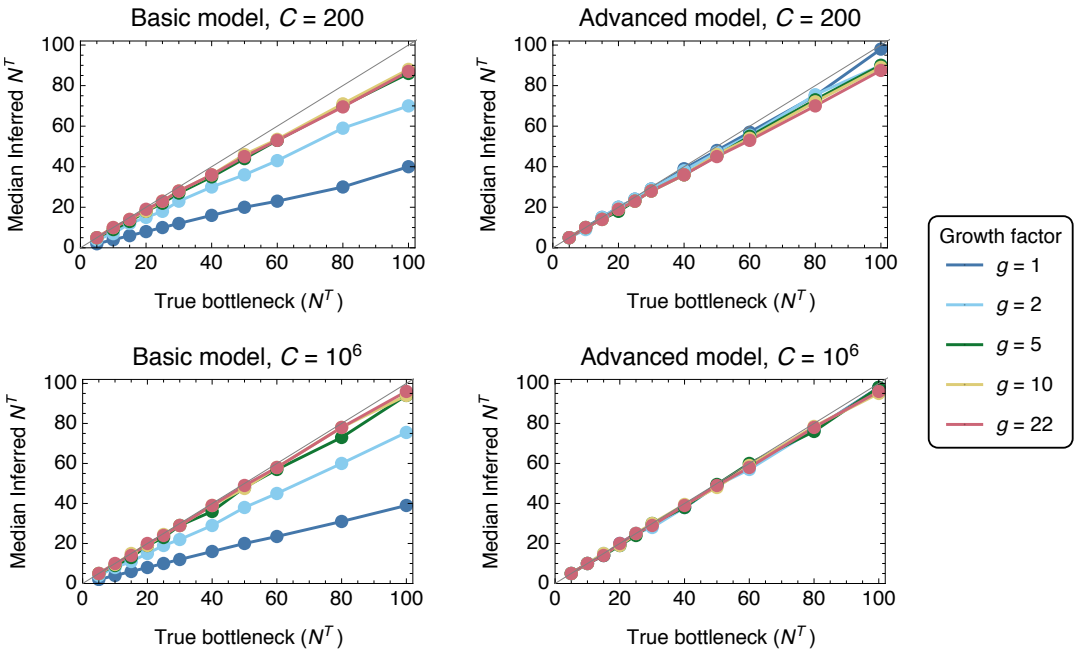


Figure 4.4. Median inferred bottleneck size from data simulating neutral transmission events with four within-host growth rounds, each exhibiting a g-fold increase in population size. Growth factors of $g = \{1, 2, 5, 10, 22\}$ and noise values of $C = \{200, 10^6\}$ were explored. Each data point represents the median bottleneck calculated over 200 replicate simulations.

the performance of the basic model to that of the advanced model which correctly accounts for the four rounds of within-host growth.

We find that the basic model significantly underestimates the bottleneck size when the growth factor is low. This can be understood as the basic model not accounting for the subsequent growth rounds which significantly alter the viral population when g is low. This effect becomes negligible as $g \geq 5$ for which the secondary growth rounds have little impact on the change in diversity compared to the initial round. Conversely, the advanced model does equally well for all values of g. We note that for our standard choice of $g = 22$ there is virtually no difference between the outcomes of the basic and advanced models. As demonstrated previously, reducing the amount of noise in the system improves our ability to infer the correct bottleneck size as N^T is increased.

4.3.1.3 Inference of Multiple Sites Under Selection

I considered the ability of the model to infer multiple instances of selection with selection acting either within a gene segment (referred to as scenario A) or on different segments (scenario B, see Methods). Figure 4.5 compares the true and

false positive rates of selection inferences of the two selection coefficients (σ^{T_1} and σ^{T_2}) in the two scenarios. True positives refer to inferences of selection upon the specific variant under selection (i.e. not both), and false positive calls describe selection inferences for neutral sites, see figure text. As expected, we observe a qualitative similarity between the two subplots for scenario A and between those for scenario B. Minor differences may be attributed to variability arising from a limited number of replicate simulations as well as discrepancies in HA and NA segment lengths resulting in differential power of inference (scenario B only). For scenario A, we observe a lower degree of true positive calls than found in the baseline case (Figure 4.2, top left). This can in part be ascribed to the method employed in determining the BIC penalty function (see Section 3.2.15) which have been tuned to identifying the first coefficient, not subsequent ones. Secondary coefficients are likely to bring about smaller improvements in BIC (as the first coefficient is optimised to capture as much of the diversity change as possible) which makes the model less able to infer both coefficients. Additionally, if the two selection coefficients act partially in the same direction it is increasingly difficult to identify individual contributions to diversity changes and to associate these with specific variants. In the limiting case where the selective pressures act in parallel (or antiparallel) the method will infer a selection coefficient of $\sigma^{T_1}+\sigma^{T_2}$ (or $\sigma^{T_1}-\sigma^{T_2}$) for one of the loci whilst neglecting the other. In other words, in the case of completely linked alleles, we have no power to distinguish between the two selection coefficients. By extension, the more polymorphisms available, the more linkage information and, in turn, the more power for separating selective effects.

Figure 4.6 displays smooth kernel distributions of the inferred selection coefficients for the true positive inferences found in Figure 4.5. Once again we observe a degree of similarity between subplots within a scenario indicating that the method is not significantly better at inferring selection for one variant than the other. The distributions in scenario A exhibit larger variances than those of scenario B, which may be attributed to the inference model conferring the accumulated selective effects of both variants onto just one of the selected alleles. This phenomenon is impossible in scenario B where the gene segments are completely unlinked.

Finally I investigated the ability of the method to jointly infer the presence of both selective effects. Figure 4.7 compares the true and false positive rates of selection inferences where true positive calls encompass outcomes for which selection was inferred at both of the selected loci. For scenario A we note a low

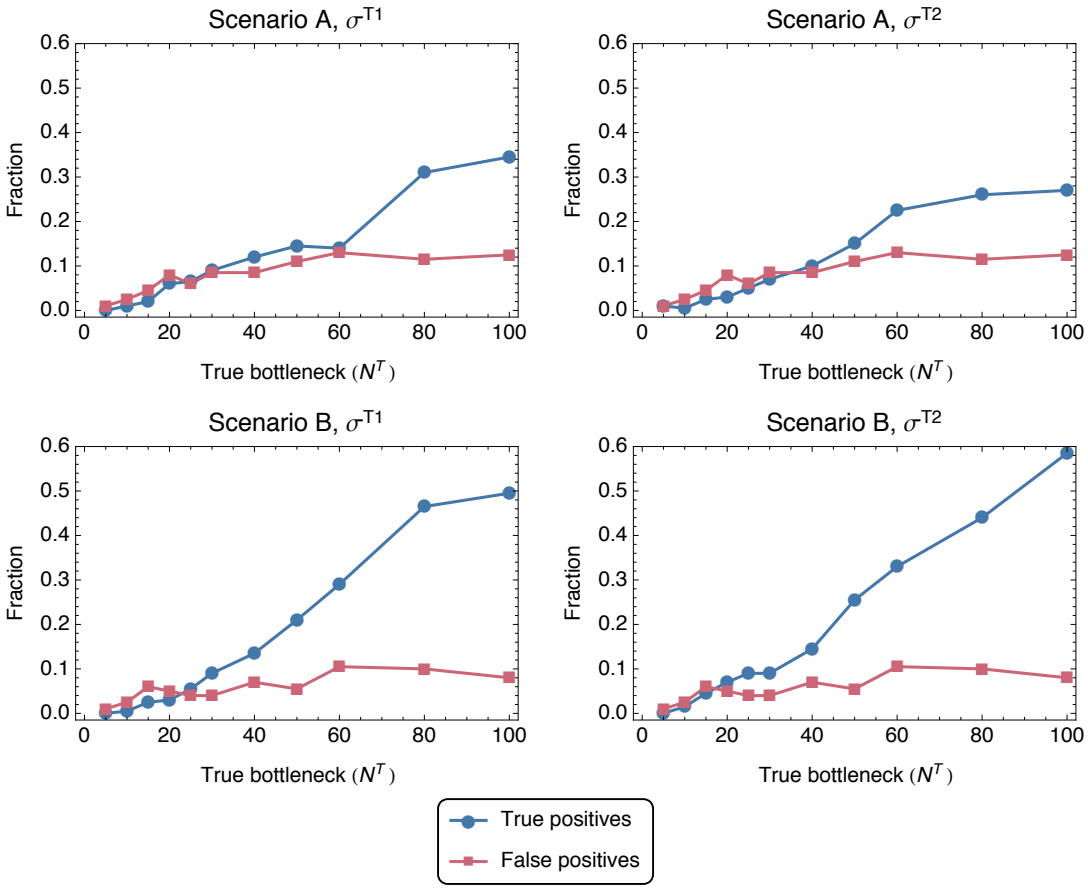


Figure 4.5. True and false positive rates of selection inference from 200 simulations of transmission events with two locus selective effects. In scenario A (top plots), selection was chosen to act on the second and fourth loci in HA with strengths $\sigma^{T_1} = \sigma^{T_2} = 1$. In scenario B (bottom plots), selection acted on the third loci in HA with strength $\sigma^{T_1} = 1$ and on the third loci in NA with strength $\sigma^{T_2} = 1$. In scenario A, true positives were defined as inferences for which selection was inferred for the second loci (top left) or the fourth loci (top right). Correspondingly, in scenario B true positives were defined as inferences for which selection was inferred for the variant in HA (bottom left) or the variant in NA (bottom right). False positives were defined as inferences for which selection was inferred at any neutral loci (i.e. excluding both of the loci genuinely under selection).

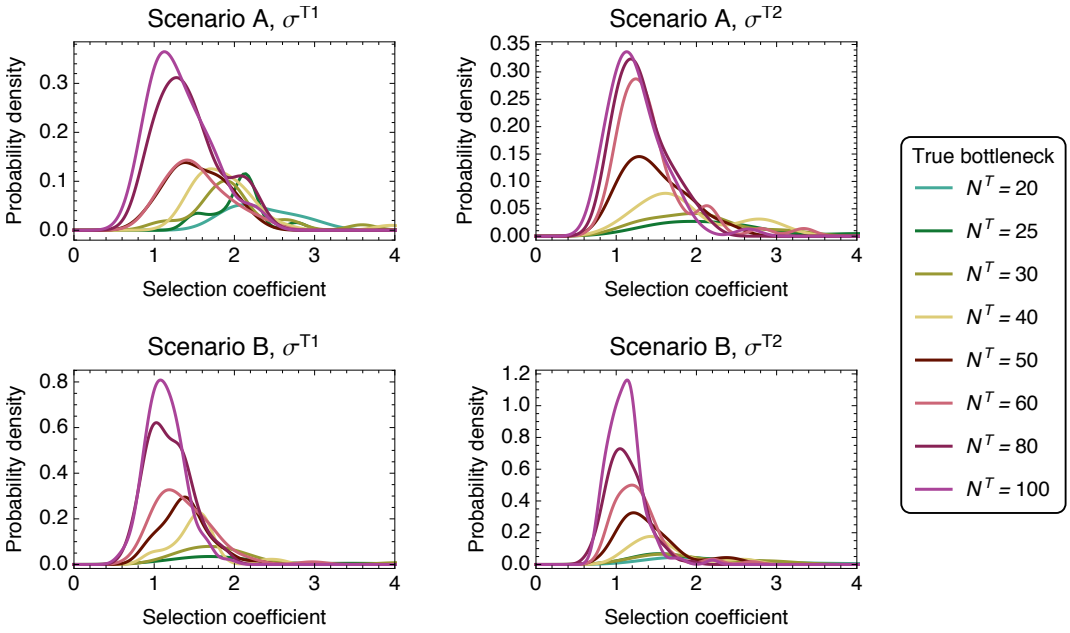


Figure 4.6. Probability distributions of inferred selection coefficients from 200 simulations of transmission events with two locus selective effects. Scenarios A and B are as described in Figure 4.5. Distributions were constructed for bottleneck values where the inference of selection resulted in a true positive rates of above 5 %. Smooth kernel distributions were computed using a Gaussian kernel function defined on $(0, 10)$ and Silverman's rule of thumb (Silverman 1986, p. 48) employed for the bandwidth size. Distributions were scaled such that their integral across the kernel range equalled the true positive rate.

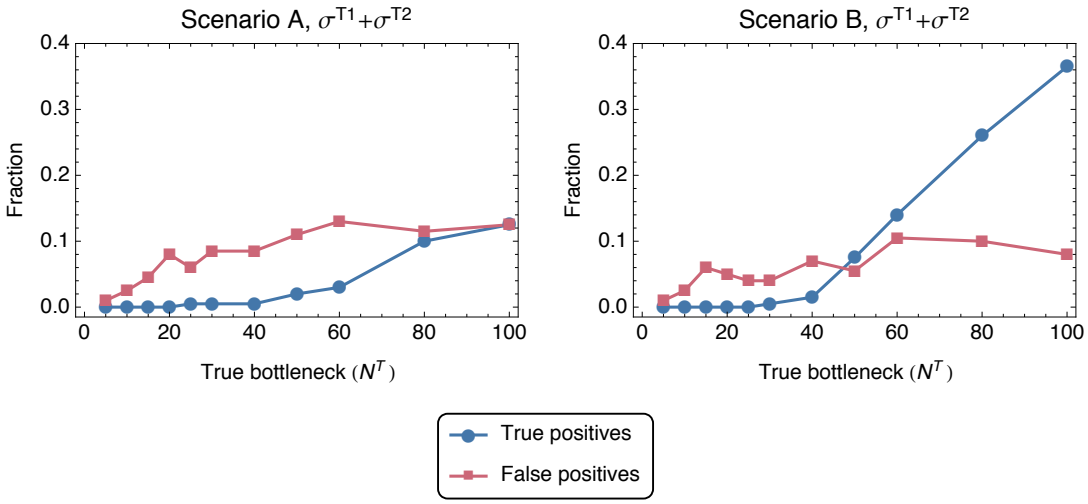


Figure 4.7. True and false positive rates of selection inference from 200 simulations of transmission events with two locus selective effects. In scenario A (left), selection was chosen to act on the second and fourth loci in HA with strengths $\sigma^{T_1} = \sigma^{T_2} = 1$. In scenario B (right), selection acted on the third loci in HA with strength $\sigma^{T_1} = 1$ and on the third loci in NA with strength $\sigma^{T_2} = 1$. True positives were defined as inferences for which selection was inferred for both of the selected loci. False positives were defined as inferences for which selection was inferred at neutral loci.

true positive rate which only approaches the false positive rate as the bottleneck reaches $N^T = 100$. On the contrary, the true positive rate in scenario B is considerably in excess of the false positive rate for $N^T \geq 50$. In the former case, the method fails to distinguish the two sources of selection, whilst in the latter case, selection acts on selectively independent alleles, leading to a significant improvement in the joint detection rate.

4.3.2 Application to an Experimental Dataset

We applied our approach to an influenza transmission dataset obtained by Watanabe et al. (Watanabe et al. 2014) and subsequently analysed by Moncla et al. (Moncla et al. 2016). This dataset provides high-resolution, whole-genome sequence data describing both the within-host evolution, and airborne transmission, of a 1918-like influenza virus, that became transmissible upon introduction of three key mutations, PB2 E627K, HA E190D and G225D. This three-mutant strain was denoted 'HA190D225D' and successfully transmitted in one of three ferret transmission pairs. Isolation and subsequent growth in MDCK cells of viruses from the contact ferret of the successful transmission led to the generation of the 'Mut' strain, which transmitted in two of three instances. A previous

analysis of these data using linked variants on the HA segment identified an increase in the diversity of the viral population during within-host growth, and respectively 'loose' and 'stringent' bottlenecks in the transmission of the two strains. In the transmission of the Mut strain, the fixation of sequence variants, potentially due to selection, was observed, while the observation of two out of three, rather than one out of three, successful transmissions suggested that the Mut virus may have evolved increased fitness for infection. Within and between hosts, segment-wide and localised measures of synonymous and non-synonymous sequence diversity π were used to assess the presence or absence of selection, leading to the conclusion that selection affected the system during transmission of the Mut strain.

In our study, data from serial samples from the within-host populations were used to infer a fitness landscape describing the within-host growth of the virus for each of the two experimental populations. Using a previously published approach (Illingworth 2015) we inferred the presence of non-neutral change in the population in seven out of eight segments in the combined HA190D225D population, and in four out of eight segments in the combined Mut population. The inference of positive selection acting for multiple non-consensus viral haplotypes in the HA segment (Figure 4.8) explains the increase in sequence diversity previously observed in these data. Further results are shown in Figures 4.9 and 4.10 and in Tables 4.1 and 4.2.

Applying our inference framework to the data identified narrow transmission bottlenecks in each case (Figure 4.11). In each of our calculations a set of statistical replicate inferences was produced, corresponding to different potential reconstructions of the population $\boldsymbol{q}^B$ from the sequence data (see Section 3.2.14.2). Within the HA190D225D population, our estimated bottlenecks ranged from 3 to 6, with a median bottleneck size of 5, while for the Mut calculations, our bottlenecks ranged from 2 to 127 and 2 to 61, with medians of 6 and 2 respectively. As such, no clear evidence was found that the HA190D225D transmission involved a greater number of particles than the Mut transmissions. Given the inclusion of the inferred within-host selection S^G, no evidence was found for the existence of variants making the virus more or less transmissible, with selection being inferred in only a small number of the replicate calculations, see Figure 4.12. Increasing the frequency cutoff at which variants were included in the calculation led to small decreases in the inferred bottleneck sizes, see Figure 4.13.

Table 4.1. Inferred fitness coefficients for HA, NA, NP, and NS for the within-host evolution of the virus within each experiment. Parameters were inferred across all index and contact ferrets within each experiment and are reported to a single decimal place. Only polymorphisms at which within-host selection was identified are listed. The parameter χ denotes an epistatic interaction between variant alleles. We note that our method infers the approximate shape of a fitness landscape based upon a reconstruction of whole viral segments; individual selection coefficients may be subject to variance between similar fitness landscapes.

Segment	Variant	Mut	HA190D220D
HA			
	G14T	0.0	0.4
	A400G	0.2	−0.3
	A507C	0.5	0.4
	C550A	0.3	-
	T634C	0.4	-
	A649G	-	0.3
	A651C	0.6	-
	T653G	0.5	-
	G741A	-	−0.1
	G747A	0.2	-
	A748G	0.3	-
	A868T	0.2	0.3
	T1036C	-	−2.1
	G1263A	-	0.5
	C1762T	−0.7	-
	$\chi_{14,400}$	−0.3	-
	$\chi_{14,507}$	0.2	-
	$\chi_{400,507}$	−36.5	-
	$\chi_{400,550}$	−32.2	-
	$\chi_{400,1036}$	-	2.2
	$\chi_{868,1263}$	-	−0.8
NA			
	G440A	-	0.4
	G649A	-	0.2
NP			
	G600A	-	0.4
NS			
	G289A	-	0.4

Table 4.2. Inferred fitness coefficients for PA, PB1, and PB2 for the within-host evolution of the virus within each experiment. Parameters were inferred across all index and contact ferrets within each experiment and are reported to a single decimal place. Only polymorphisms at which within-host selection was identified are listed. The parameter χ denotes an epistatic interaction between variant alleles. We note that our method infers the approximate shape of a fitness landscape based upon a reconstruction of whole viral segments; individual selection coefficients may be subject to variance between similar fitness landscapes.

Segment	Variant	Mut	HA190D220D
PA			
	A781G	−0.2	-
	G1500T	-	0.5
	C1651T	−0.7	-
	G1880T	−0.8	-
PB1			
	C65T	-	−0.4
	C90A	−0.3	−0.3
	C835A	-	0.3
	G982T	-	0.4
	T1151G	-	0.3
	G2250T	−1.0	-
	$\chi_{90,982}$	-	1.0
PB2			
	A1199G	−0.9	-
	T1537C	-	0.5
	G2193C	-	−0.3

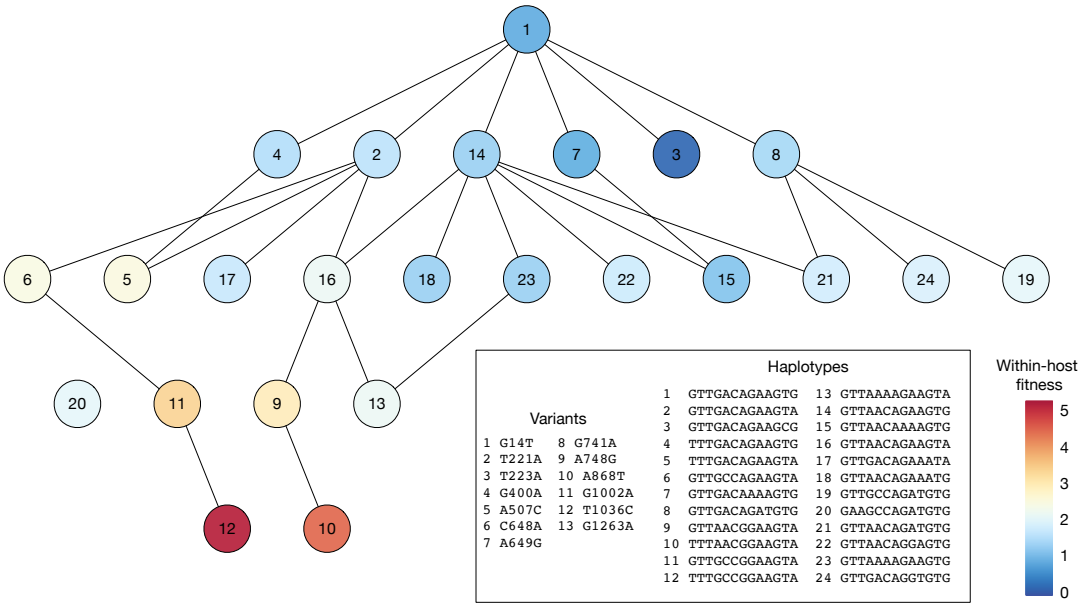


Figure 4.8. Inferred fitness landscape for within-host growth using data from the HA190D225D dataset. Viral haplotypes for which the inferred frequency rose above 1% in at least one animal are shown. Lines show haplotypes separated by a single mutation.

4.4 Discussion

In this chapter I defined the advanced transmission model and used this to investigate high complexity transmission scenarios. The advanced model attempts to distinguish between selection for increased transmissibility and selection for within-host adaptation. Assuming an accurate estimation of within-host infection, the model was able to account for this selection by defining a baseline amount of genetic change such that additional diversity modifications due to selection for transmission were immediately apparent. When not accounted for, selection for within-host adaptation was misclassified as selection for transmission. This demonstrates that within-host selection is yet another confounding factor that needs accounted for if a proper inference of transmission is to be achieved. Whilst not explicitly shown, it is worth noting that incorrectly accounting for within-host selection doesn't just obscure selection inference, but also bottleneck inference when utilising a neutral inference method. I note that my method is able to account for within-host selection and infer a neutral bottleneck with respect to selection for transmission. Given an estimate of within-host selection, the neutral model is computationally efficient and doesn't require the inference of BIC penalty functions. As alluded to previously, when a dataset consists of just two time points it is mathematically impossible to distinguish different

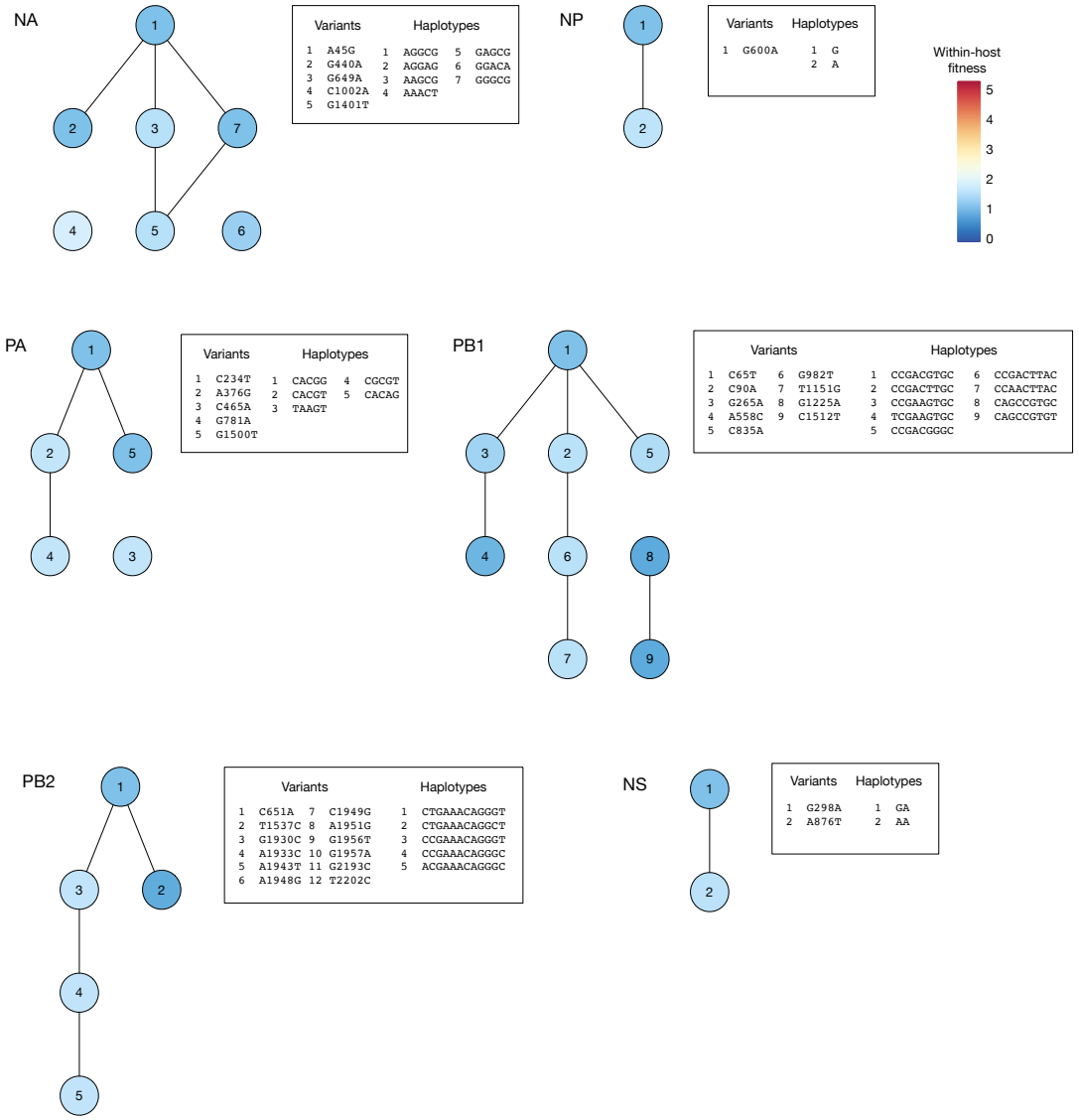


Figure 4.9. Inferred within-host fitness landscape for segments in the HA190D220D viral populations. Haplotypes for which the inferred frequency rose to a frequency of at least 1% in at least one animal are shown. Haplotypes which are separated by a single mutation are joined by lines.

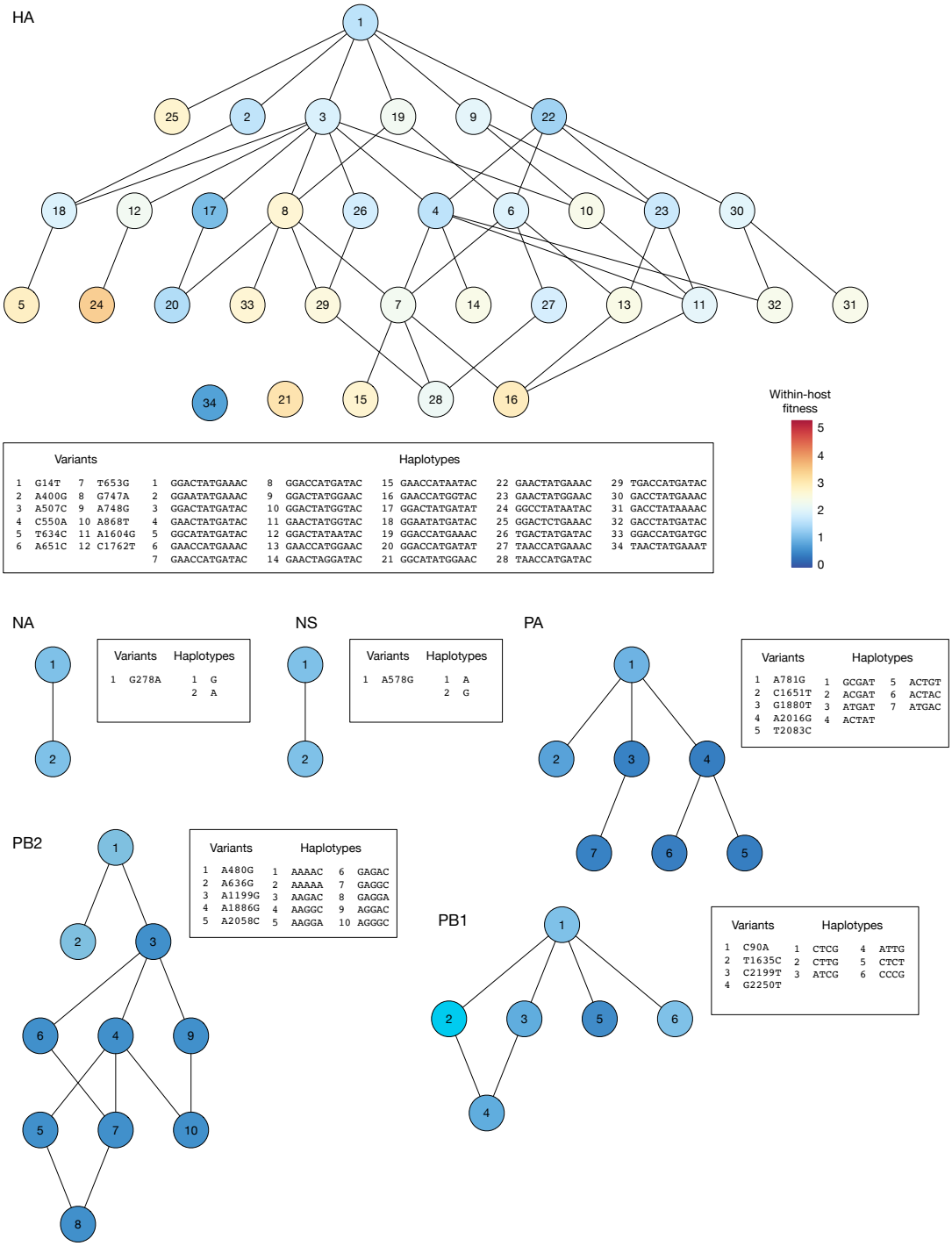


Figure 4.10. Inferred within-host fitness landscape for segments in the Mut viral populations. Haplotypes for which the inferred frequency rose to a frequency of at least 1% in at least one animal are shown. Haplotypes which are separated by a single mutation are joined by lines.

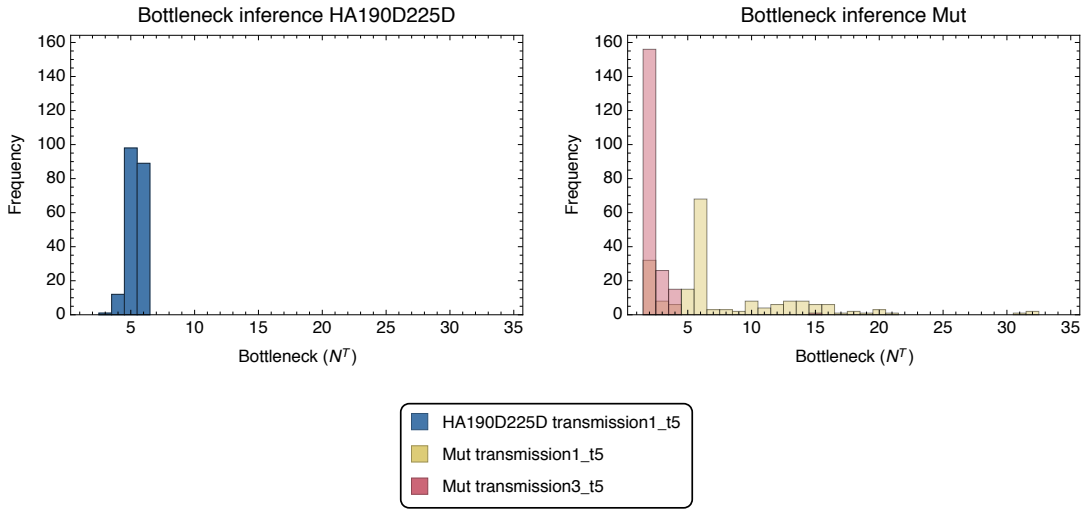


Figure 4.11. Histograms of bottleneck inferences for HA190D225D and Mut transmission pairs from 200 analysis seeds. A replicate inference method was employed for the Mut transmission pairs such that a common fitness landscape was imposed. The Mut transmission pairs may take different bottleneck values and have been plotted as an overlapping histogram. Bottleneck inferences larger than $N^T = 35$ have been omitted for clarity.

sources of selection. Considering the importance of accounting for within-host selection, I propose that future transmission studies adhere to a frequent and regular sampling schedule, allowing for the characterisation of the within-host fitness landscape of the viral population.

I have here made the modelling assumption that within-host growth is characterised by two distinct processes, namely a stochastic increase in population size followed by modifications due to selection. Within-host selection covers adaptation across multiple different aspects of the growth process, including viral transport to the cell nucleus, viral replication, formation of new viruses, budding of virions, and infection of new host cells. I here argue that within-host adaptation primarily occurs during or after the replication process, thereby supporting the specific ordering wherein replication comes before selection. This is a model choice; selection obviously takes place throughout the growth process, but, given limited resolution to distinguish the effects further, I here make the assumption that selection has the most effect during later stages of proliferation.

Considering more realistic within-host growth processes consisting of multiple rounds of viral replication, the basic transmission model was found to underestimate the bottleneck size when considering small growth factors. This may be understood as the basic model not accounting for the increase in viral

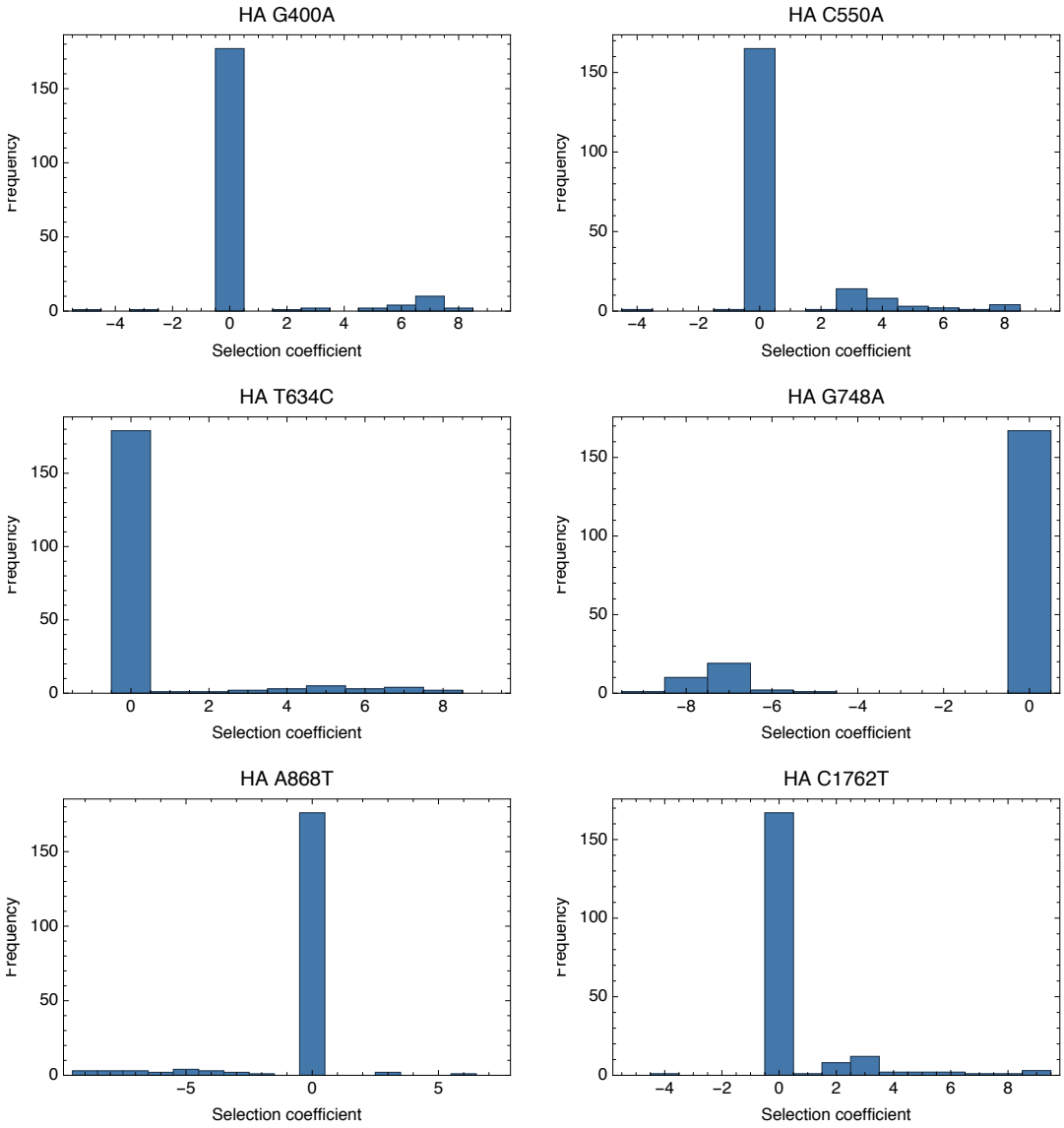


Figure 4.12. Histograms of selection inferences for the Mut transmission pairs from 200 seeds using an allele frequency cut-off of 2%. A replicate inference method was employed such that a common fitness landscape was imposed. Selection inferences that resulted in at least 10% non-zero inferences are here reported by the nucleotide position of the variant site.

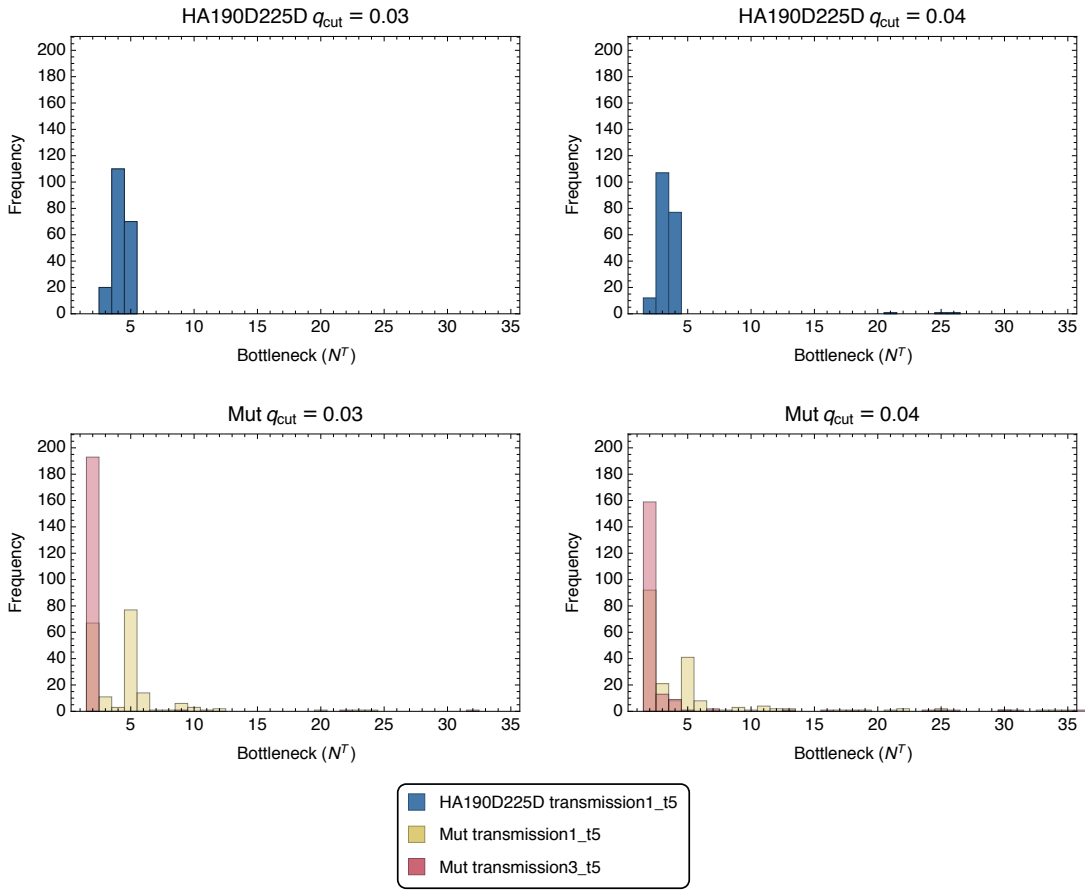


Figure 4.13. Histograms of bottleneck inferences for HA190D225D and Mut transmission pairs from 200 random seeds using allele frequency cut-offs of $q_{\text{cut}} \in \{0.03, 0.04\}$. A replicate inference method was employed for the Mut transmission pairs such that a common fitness landscape was imposed. The Mut transmission pairs may take different bottleneck values and have been plotted as overlapping histograms. Bottleneck inferences larger than $N^T = 35$ have been omitted for clarity.

diversity arising from subsequent growth rounds. On the contrary, the advanced model performed equally well for all examined scenarios. For growth factors of 5 or more, the inferences of the basic model coincided with those of the advanced model. These results suggest a theoretical advantage in accounting for multiple growth rounds, however, in practise this is irrelevant for more realistic growth factor values. For simulations in general and for the analysis of real datasets I have assumed a 22-fold growth process. In this instance it is sufficient to account for a single 'effective' round of within-host growth.

As a final evaluation of the transmission model I considered the ability of the method to infer multiple instances of selection. The method substantially underperformed when inferring multiple selection coefficients acting within a single gene segment. Here the method was more likely to infer a single site under selection instead of two. A number of possible factors may contribute to this. Firstly, the selection inference framework was optimised to find a single instance of selection, not multiple. In theory, this could be corrected by tuning the BIC penalty differently, e.g. by having different BIC penalty curves for each additionally inferred parameter. Secondly, linked alleles may result in the method partially attributing selection acting on one allele to the other. In the limit of completely linked sites, distinguishing which allele is under selection becomes impossible. In cases where selection acts on different segments there is scope for inferring multiple selection coefficients, at least under the assumption of rapid reassortment.

Separating selection from drift is an inherently difficult task. Accounting for multiple instances of selection is increasingly difficult and generally only feasible when specific biological considerations apply. I conclude that my method is capable of inferring single selection coefficients when selection is strong and stochastic effects small, but that further work on the comparison of likelihood statistics is required in order to correctly derive more complicated selection models. Whilst potentially possible, an extension of the adaptive BIC framework to account for more complicated aspects of selection represents a substantial step away from the BIC scheme and a step towards a more machine learning oriented approach. Such approach is out of the scope of this work and will not be considered here.

Considering our analysis of data from a recent evolutionary experiment, our approach provides a greater precision in the inference of evolutionary statistics, leading to an alternative explanation for the data observed. Where data have previously been interpreted as implying differential transmission bottle-

necks between strains, our approach infers bottlenecks of similar sizes ranging from 2-6 viruses. Furthermore, where evidence has been interpreted to suggest a differing extent of transmissibility between strains, our approach attributes changes in the composition of the population to a mixture of stochastic effects and selection for increased within-host adaptation. Our result does not prove the absence of differential transmissibility among the viruses involved in this study; at the bottleneck size we inferred, selection is very hard to identify even where it does influence transmission. Rather, our claim is that under a parsimonious analysis of the data, apparent evidence for increased viral transmissibility can be explained by other evolutionary factors.

Chapter 5

Analysis of Experimental Study on Influenza Transmission in Pigs

5.1 Introduction

In the previous chapter I extended the basic transmission model to account for a more detailed description of within-host growth processes. I considered the inclusion of selection for increased within-host adaptation and derived a general solution for a multi-step within-host growth process. It was observed that improperly accounting for within-host selection led to a bias in the inference of selection for transmissibility.

In Chapters 3 and 4 my analysis was based upon the assumption that viral populations underwent a high rate of viral reassortment. Such an assumption is supported by experimental evidence from *in vitro* and small animal systems, and makes the calculations of selection and bottleneck sizes more computationally tractable. In this and the following chapter I note that the assumption of rapid reassortment may not always hold; in large mammalian systems the experimental evidence does not support this assumption. In such conditions the framework I have set out above may not be computationally feasible; for example computing within-host fitness landscapes becomes more difficult as more polymorphic loci exist in mutual linkage disequilibrium. I therefore examine approximations to the above framework; a full accounting for selection is not always achievable, but steps may be taken towards a best approximate solution, which seeks to improve upon a model of complete selective neutrality.

In this chapter I consider the application of my model to an influenza transmission study in small herds of swine. Discussing the assumptions and limita-

tions associated with selection inference, I define an effective selection framework accounting for within-host selection in hosts exhibiting low levels of viral reassortment. After applying this method to experimental sequencing data, I discuss the ability of consensus and sub-consensus methods for inferring transmission networks in relation to the dataset at hand.

5.1.1 Effective Selection

The model presented so far describes a multi-locus approach to transmission inference. Adopting a multi-locus framework allows us to identify the presence of selection within single gene segments, with linkage disequilibrium information being obtained through the use of partial haplotypes. By contrast, single-locus models fail to capture the relationship between variant alleles and as a result make it impossible to determine the effect of selection upon individual loci. By extension, separating selective effects between different viral segments, where linkage disequilibrium between alleles on segments may exist, requires a multi-segment, multi-locus haplotype model. Such a model has been successfully used by Sobel Leonard et al. (2017a) to investigate the extent of reassortment of influenza virus in human hosts, however, it results in a substantial computational overhead when the number of loci (and in turn haplotypes) is large; the lack of cross-segment information in short-read data means that the number of haplotypes in the model is equal to the product of the number of haplotypes in each viral segment. The data analysed by Sobel Leonard et al. had polymorphisms in only four out of the eight viral segments, and in general, such multi-segment approaches are only applicable in cases where only a handful of variants are present. Steps to simplify the model may be required for a calculation to be computationally possible.

In my previous analyses I made the assumption of infinite reassortment of genes during viral replication. This assumption guarantees the removal of across-segment linkage effects and is a key requirement for separating the contributions of drift and selection, see Figure 3.4. This assumption is believed to hold for *in vitro* and small animal studies (Ince et al. 2013; Marshall et al. 2013; Tao, Steel and Lowen 2014; Tao et al. 2015), but it has recently been shown that the effective reassortment rate is highly limited in human influenza infections (Sobel Leonard et al. 2017a). In this and the following chapter I discuss possible approaches which allow us to sidestep these restrictions by applying an *effective within-host selection* framework.

5.1.2 Swine Flu and Emergence of Pandemics

In addition to humans and birds, the natural host of flu, influenza viruses infect a range of other mammals such as pigs, horses and dogs (Crawford et al. 2005; Murcia et al. 2010; Webster et al. 1992). Existing in large herds and being in frequent contact with humans, pigs represent a potential risk of animal to human transmission of influenza virus. In fact, transmission from pigs to humans and from humans to pigs have been extensively observed (Ma, Kahn and Richt 2008; Shortridge et al. 1977). Harbouring both SAα2,3Gal receptors, to which avian influenza viruses preferentially bind, and SAα2,6Gal receptors, to which human influenza viruses primarily attach, pigs are susceptible to infection by both human and avian influenza viruses (Kawaoka et al. 1998). As a result, pigs have been suggested to act as a 'mixing vessel', where, if co-infected by human and avian influenza strains, reassortment of gene segments might lead to the creation of novel, highly pathogenic, viral strains (Haß et al. 2011; Ma, Kahn and Richt 2008; Shortridge et al. 1977). Studies have considered the adaptation of potentially reassortant, avian-derived viruses for transmission in mammals, suggesting that as few as a handful of mutations may lead to efficient transmission in the ferret model (Herfst et al. 2012; Imai et al. 2012; Watanabe et al. 2014; Wilker et al. 2013).

In the past one hundred years, four influenza pandemics have had devastating impacts upon humanity. These pandemic strains derived, at least partially, from animal reservoirs, with the latter three (1957, 1968, 2009) having been confirmed as arising following reassortment events (Garten et al. 2009; Ma, Kahn and Richt 2008; G. J. D. Smith et al. 2009). The recent 2009 H1N1 swine flu pandemic spread to more than 214 countries worldwide and officially claimed more than 18,000 lives (World Health Organization 2010). This pandemic arose from reassortment events in pigs with the NA and M genes deriving from the Eurasian swine linage and the remaining gene segments exhibiting identity to those of the triple-reassortant swine lineage (Garten et al. 2009; G. J. D. Smith et al. 2009). Alarmingly, segments with the highest similarity to the pandemic strain were on average isolated ten years prior to the outbreak, suggesting that the ancestral strains had been circulating undetected for a significant period of time.

With a view to determining intra- and inter-host evolution, Murcia et al. (2012) analysed transmission events from Eurasian avian-like swine influenza virus of strain A/swine/England/453/2006 (H1N1) in chains of naïve and vac-

cinated pigs. Considering variants in HA obtained from capillary sequencing of nasal swab samples, a loose transmission bottleneck was hypothesised on the basis of identification of multiple variants persisting across transmission events. An analogous approach was applied in a horse transmission study with equine influenza virus of subtype H3N8 for which a similar conclusion was reached (Murcia et al. 2010). Additionally it was shown that the mutations required for the establishment of infection in dogs were present in horse populations, suggesting that the equine-to-canine spillover event at the start of the millennium (Crawford et al. 2005) may have readily occurred and that adaptation to a new species may take place in the original host prior to transmission.

5.1.3 Inference of Transmission Networks

Improving methods for solving route of transmission problems is of importance for determining sources of infections and for developing preventive approaches to the spread of pathogens. Traditionally, who-infected-whom problems have been considered at the consensus sequence level, providing insight only where viral populations have consensus sequence differences; HIV infections, in which genetic variation accumulates over several years, provide an example where this can be achieved (Leitner et al. 1996). Whilst informative, these approaches are disadvantaged when considering acute infections for which the number of consensus sequence differences may be minimal or absent.

The increasing application of next-generation sequencing to viral outbreaks has been matched by the development of methods that identify transmission networks on the basis of sub-consensus measures. The enhanced resolution inherent to short-read data allows for the evaluation of within-host diversity statistics, considering for instance the amount of shared diversity between hosts (Worby, Lipsitch and Hanage 2017). Where two viral populations have a large degree of shared diversity, this is interpreted as evidence in favour of a potential transmission link. Worby et al. notes that where a transmission event is governed by a bottleneck of a single virion, any subsequent diversity observed in the recipient must have arisen entirely *de novo*.

Worby et al. carried out an in-depth analysis of the performance of consensus and shared variant methods for the inference of transmission networks. Attempting to reconstruct patterns of transmission from simulated outbreaks, three methods were employed for assignment of transmission links: 1) A maximum (shared) variant tree, 2) a minimum (genetic) distance tree, and 3) a

hybrid maximum tree. Where shared polymorphisms are present, the maximum variant approach identifies transmission hosts as those harbouring the largest number of shared variants with the recipient in question. The minimum distance approach determines host-recipient links by minimising the Hamming distance between consensus sequences of potential transmission pairs. Finally, the hybrid maximum method combines the above approaches, aiming firstly to identify transmission sources by maximising the number of shared variants, and then, if no shared variants exists, resorting to the minimum distance approach. From simulations, the maximum variant approach was found to be effective at bottleneck sizes of eight or larger, with hybrid and minimum distance approaches outperforming it for narrow transmission bottlenecks. In the case of small bottlenecks, a large degree of shared polymorphisms are likely to be lost, resulting in minimum distance metrics becoming relatively more informative. Considering instead the true path distance between estimated transmission pairs, the maximum variant approach was found to outcompete the minimum distance method across a range of mutation rates, suggesting that the conclusions of the study are somewhat robust and applicable to outbreaks from different infectious diseases.

Employing a shared variant scheme, Stack et al. (Stack et al. 2013) attempted to infer transmission networks in the horse (Murcia et al. 2010) and pig (Murcia et al. 2012) studies described above. On the basis of an initial network, a refined explanation was invented by employing a classification scheme utilising known constraints, i.e. by computing importance scores for animals potentially in contact. This process removed the majority of inconsistent edges in the network, however, it also drastically reduced the number of consistent edges. The authors did not comment on whether the method is potentially too conservative in restricting the size of the network. Importantly, consensus sequence methods were deemed ineffectual on the basis of a limited number of fixations events in the studies.

5.1.4 Author Contributions

The work presented in this chapter is currently unpublished. The work described here was carried out by the author under the supervision of his PhD supervisor, Dr Christopher Illingworth. The within-host selection inference code was written by Chris Illingworth for a previous publication (Illingworth 2015). The swine transmission study analysed in this chapter was collected in July 2010 and is

currently unpublished. We kindly thank the following people for allowing us to access the data prior to publication: S. Brookes[1], A. Germundsson[1], C. Inman[2], M. Bailey[2], S. Dunham[3], G. White[3], F. Garcon[1], A. Núñez[1], R. Saenz[4], J. Gog[4], T. Freeman[5], R. Kapetanovic[5], A. Tomoiu[5], C. Donnelly[6], K-C. Chang[3], A. Archibald[5], COSI[7], J. Wood[8], I. Brown[1].

5.2 Methods

5.2.1 Effective Within-Host Selection

The fitness of a virus describes the replicative ability of the virus as a function of its genotype and the specific host system it resides within (Sobel Leonard et al. 2017a). Collecting fitnesses for all potential haplotypes, one can construct a fitness landscape which is constant in time under the assumption that the host environment doesn't change. For the Moncla et al. dataset we assumed a high rate of reassortment, effectively yielding the fitness of an individual segment independent from the fitnesses of the remaining segments. Within-host selection was estimated in terms of an underlying fitness landscape for each segment. In the schema we use, this landscape is built up of a sum of single-locus and two-locus fitness components; we use a process of model selection to derive the most parsimonious explanation for the data under a maximum likelihood framework, see Section 4.2.7.

Whilst an elevated rate of reassortment may be assumed in small animal studies, this assumption is potentially invalid for larger mammals for which linkage effects between viral segments may influence the within-host evolution of individual alleles (Sobel Leonard et al. 2017a). Theoretically, a constant fitness

[1]Animal Health and Veterinary Laboratories Agency-Weybridge, EU/OIE/FAO Reference Laboratory for Avian Influenza and Newcastle Disease, Addlestone, Surrey, UK. KT15 3NB

[2]School of Clinical Veterinary Science, University of Bristol, Langford House, Langford, Bristol, UK. BS40 5DU

[3]School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonington Campus, College Road, Loughborough, Leicestershire, UK. LE12 5RD

[4]Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge, UK. CB3 0WA

[5]Division of Genetics and Genomics, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian UK. EH25 9RG

[6]MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, Faculty of Medicine, Imperial College London, St Mary's Campus, Norfolk Place, London, UK. W2 1PG

[7]Combating Swine Influenza, Wellcome Trust-MRC-BBSRC-Defra UK consortium

[8]University of Cambridge, Department of Veterinary Medicine, Madingley Road,Cambridge, UK. CB3 0ES

landscape may be constructed by accounting for multi-segment haplotypes and by associating reassortment rates describing how fitness effects from specific segments interfere with each other. However, this is computationally demanding in all but the simplest of cases. For a single segment we may have e.g. five variant loci resulting in $2^5 = 32$ potential haplotypes. In the case of a multi-segment framework, the number of potential haplotypes is the product of the number of segment-specific haplotypes taken across all segments; potentially an extremely large number. In this circumstance the inference of a fitness landscape is impractical. As a compromise, we propose an alternative means of estimating the within-host fitness for the purposes of our model. The fitness of a variant can be decomposed into the sum of its own inherent fitness, which describes the inherent advantage or disadvantage the variant contributes to the virus, plus an interference term, which describes the influence of other variants, in linkage disequilibrium with the first variant, upon the variant itself. The sum of the inherent fitness plus interference can be termed the 'effective' fitness of the variant (Illingworth and Mustonen 2011).

Based upon this principle, we attempt to derive an *effective fitness landscape* for each segment, capturing the combination of inherent fitness plus interference effects.

Our interest is in the manner in which within-host selection affects the virus during the very first stage of its growth within the host. We therefore consider data from the first two time points for which data is collected from the recipient animal. Taking these two samples, we conduct an inference of within-host selection for each segment. Although the inferred fitness landscape is a composite of inherent selection and interference, it provides an estimation of how selection as a whole shapes the within-host population during the early stages of viral growth.

Our use of only two samples reflects an assumption about interference. The selective effects of interference depend upon linkage disequilibrium between alleles, which itself depends upon the composition of the population. As the genetic composition of the population changes over time, interference, and therefore the effective selection itself, is time-dependent. In so far as we are interested in how within-host selection affects the virus during the very first stage of its growth within the host, we consider data from the first two time points for which data is collected from the recipient animal, carrying out our inference under the assumption that selection acts in a constant manner for this period of time. The result is an approximation of the effect of within-host selection in

a circumstance where this would not otherwise be possible to calculate. As the effective selection depends on the genetic composition of the population, fitness landscapes are defined for each recipient animal individually, rather than across all recipients as was the case for the Moncla et al. dataset.

5.2.2 Determining Route of Transmission From Bottleneck Inference

Considering the evidence for shared variants informing the structure of transmission networks, we hypothesised that our transmission inference method, which is based on changes in diversity, might provide similar insights into route of transmission problems. In fact, our method considers both shared variants and variants unique to the host, attempting to parsimoniously explain an observed outcome. For instance, where variants are observed only in the host, the neutral version of our model infers a narrow bottleneck, attributing the lack of shared variants to a founder effect. Conversely, variants unique to the recipient are assumed to have arisen *de novo* and are ignored (see Section 3.2.13). To this end we proposed that where multiple potential transmission events exist, the more likely transmission link is that resulting in the largest inference of bottleneck size. Our method is potentially more sensitive than the maximum variant method of Worby et al. (2017) as it considers specific changes in diversity rather than simply the absolute number of shared polymorphisms. Regardless, similarly to the maximum variant method, our method has low inference power for narrow bottlenecks for which there is little resolution to distinguish true and false transmission links.

5.2.3 Determining Route of Transmission From Sub-Consensus Sequence Distance Metric

As an alternative to route of transmission inference based on traditional consensus level minimum distance methods, we proposed a sub-consensus sequence distance metric accounting for the subtle nuances arising from short-term evolution. The metric was defined as

$$\Theta = 0.5 \sum_{i} \sum_{j \in \{\mathrm{A,C,G,T}\}} |q^{A}_{i,j} - q^{B}_{i,j}| \tag{5.1}$$

where the sum i is over all sites in the genome and $|\cdot|$ denotes absolute value. The term $q^A_{i,j} - q^B_{i,j}$ describes the difference in frequencies of the jth allele at the ith site between the after (A) and before (B) populations. The prefactor of 0.5 ensures that changes of an entire unit of frequency contributes exactly unity to the metric. For instance, if the before frequencies at the ith site were $\boldsymbol{q}^B_i = \{q^B_{i,\mathrm{A}}, q^B_{i,\mathrm{C}}, q^B_{i,\mathrm{G}}, q^B_{i,\mathrm{T}}\} = \{1,0,0,0\}$ and the after frequencies $\boldsymbol{q}^A_i = \{q^A_{i,\mathrm{A}}, q^A_{i,\mathrm{C}}, q^A_{i,\mathrm{G}}, q^A_{i,\mathrm{T}}\} = \{0,0,1,0\}$, i.e. a change from allele A to G, the metric for this site would be

$$\Theta_i = 0.5 \sum_{j\in\{\mathrm{A,C,G,T}\}} |q^A_{i,j} - q^B_{i,j}| = 0.5\left(|0-1| + |0-0| + |1-0| + |0-0|\right) = 1 \tag{5.2}$$

Accounting for genome size, we may define a normalised metric:

$$\theta = \frac{\Theta}{\#\text{ of sites}} \tag{5.3}$$

The smaller the sub-consensus sequence distance metric, the more similar the before and after samples are. Comparing multiple potential transmission events, the transmission with the smallest value of θ represents the most likely true transmission.

5.2.4 Transmission Study in Pigs

As an example of transmission inference under effective within-host selection I considered a transmission study in pigs, originally obtained by Brookes et al. (2010) at the Department of Veterinary Medicine, University of Cambridge. The study examined here was part of a larger influenza study in pigs consisting of three parts, the sub-experiments denoted by A1, A2 and A3. I here consider only the A2 study and will hereafter refer to it simply as 'the study'.

5.2.4.1 Outline of Study

A schematics of the study design is shown in Figure 5.1. In total, the study involved 22 immunologically naïve pigs aged 12 weeks. Initially, four pigs (denoted 4N) were inoculated with influenza virus of the pandemic strain A/England/195/2009 (H1N1). At two days post inoculation (dpi), six additional pigs (denoted 6N-c1) were introduced to the four infected pigs. After two days of contact (4 dpi), the six 6N-c1 pigs were removed from the enclosure and replaced by six

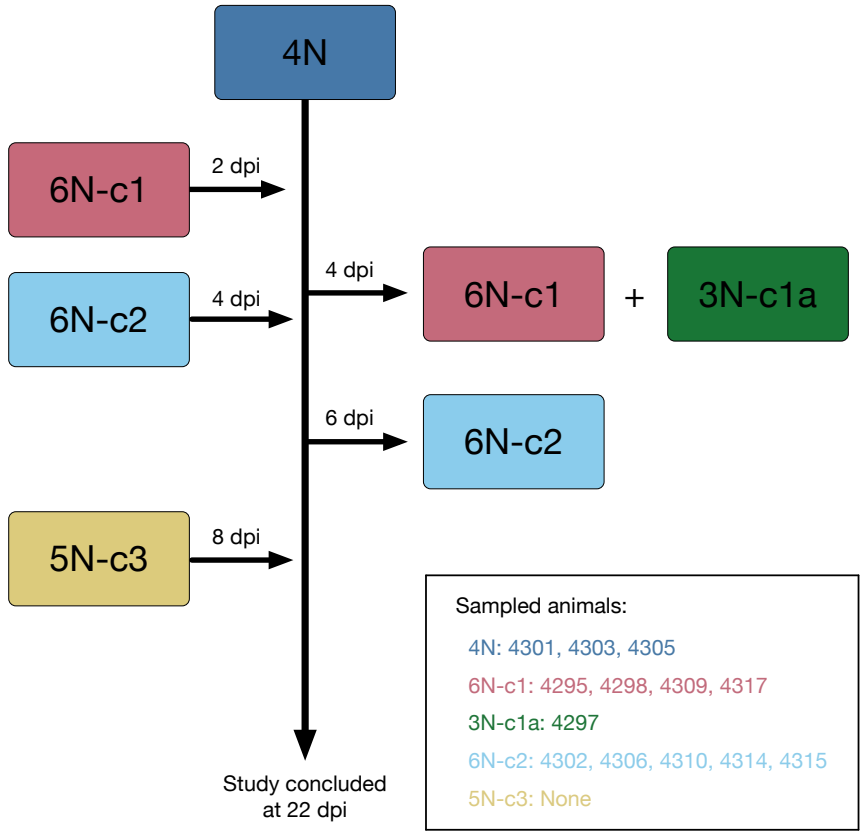


Figure 5.1. Overview of the study design used in the pig transmission study. Four naïve pigs (4N) were inoculated with influenza virus and introduced to six pigs (6N-c1) at two days post inoculation (dpi). At four dpi the 6N-c1 pigs were removed and placed in contact with three new pigs, 3N-c1a. Simultaneously, six new pigs (6N-c2) were introduced to the original 4N pigs and subsequently removed two days later. Finally, five naïve pigs (5N-c3) were introduced to the 4N pigs and the study was terminated after 22 days. Sequence data is available for a subset of these animals, see inset.

new pigs, denoted 6N-c2. The original six 6N-c1 where placed in contact with three naïve animals (denoted 3N-c1a). After an additional two days (6 dpi), the second group of six pigs (6N-c2) were removed from the enclosure with the original four pigs (4N). At eight days post inoculation, a final group of five naïve pigs were introduced to the original 4N pigs. The study was concluded after 22 days.

Samples were collected for multiple time points and paired-end read sequence data obtained for specific cases. An overview of the available samples are shown in Table 5.1.

Table 5.1. Sampling times for animals in pig transmission study. Times are with respect to inoculation (time=0) of the 4N pigs.

Time	4N			6N-c1				3N-c1a	6N-c2				
	4301	4303	4305	4295	4298	4309	4317	4297	4302	4306	4310	4314	4315
0	·	·	·	·	·	·	·	·	·	·	·	·	·
1	·	·	·	·	·	·	·	·	·	·	·	·	·
2	×	×	×	·	·	·	·	·	·	·	·	·	·
3	×	×	×	·	·	·	·	·	·	·	·	·	·
4	×	×	×	·	·	·	·	·	·	·	·	·	·
5	·	·	×	·	·	·	·	·	·	·	·	·	·
6	·	·	·	·	·	×	×	·	·	·	·	·	·
7	·	·	·	×	×	×	×	·	·	·	·	·	·
8	·	·	·	×	×	×	·	·	·	×	·	×	·
9	·	·	·	·	×	·	·	·	×	×	×	×	×
10	·	·	·	·	·	·	·	×	×	·	·	·	×
11	·	·	·	·	·	·	·	×	·	·	·	·	×

Table 5.2. Time points for the 12 potential transmission events involving the four seeder pigs (4N) and the first set of recipient animals (6N-c1). Time points are with respect to the time of inoculation of the 4N pigs (t=0).

Recipient	4301		4303		4305	
	Before	After	Before	After	Before	After
4295	4	7	4	7	5	7
4298	4	7	4	7	5	7
4309	4	6	4	6	5	6
4317	4	6	4	6	5	6

Table 5.3. Time points for the four potential transmission events involving the first set of recipient animals (6N-c1) and the secondary set of contact animals (3N-c1a). Time points are with respect to the time of inoculation of the 4N pigs (t=0).

Recipient	4295		4298		4309		4317	
	Before	After	Before	After	Before	After	Before	After
4297	8	10	9	10	8	10	7	10

5.2.4.2 Potential Transmission Events

Our transmission inference framework requires designated host and recipient populations, i.e. knowing that individual X infected individual Y is critical to using the method. For the Brookes et al. study we lack this information, knowing only that e.g. any of the four seeder pigs (4N) could have infected any of the six recipient pigs (6N-c1). Given the available data and assuming no within-group transmission we may construct 28 potential transmission events as seen in Tables 5.2 to 5.4. Transmissions involving recipient animal 4315 were ignored due to low coverage.

Table 5.4. Time points for the 12 potential transmission events involving the four seeder pigs (4N) and the second set of recipient animals (6N-c2). Time points are with respect to the time of inoculation of the 4N pigs (t=0).

Recipient	4301		4303		4305	
	Before	After	Before	After	Before	After
4302	4	9	4	9	5	9
4306	4	8	4	8	5	8
4310	4	9	4	9	5	9
4314	4	8	4	8	5	8

5.2.4.3 Data Processing

Data processing for the pig transmission study was split into four steps: 1) Initial processing using SAMtools (H. Li et al. 2009) and SAMFIRE (Illingworth 2015, 2016), 2) inference of C-values, 3) processing for within-host selection inference, 4) processing for transmission inference. Scripts for data processing can be found in the online repository, see Section 3.2.17.

Initial Processing Firstly, .sam files were extracted from .bam sources using SAMtools (H. Li et al. 2009). These files were then filtered using SAMFIRE (Illingworth 2015, 2016) and single-locus trajectories were computed at a frequency cut-off of 0.02. The `--repq 2` flag was invoked, designating that a polymorphism must be observed in at least two time points to be included in a trajectory. This step filters out false positive calls of polymorphic sites which may arise from noise in the sequencing process.

Inference of C-Values As with the Moncla et al. dataset, two noise parameters were inferred: A standard ($C = 186.33$) and a conservative ($C = 60.08$) value. The standard value was computed on the basis of the single-locus trajectories described above. To avoid a biased inference, I here removed duplicate trajectories (defined as trajectories covering the same loci, i.e. multiallelic loci), retaining the trajectory with the largest average polymorphism. The noise inference was then computed using the SAMFIRE command `sl_noise` on the basis of trajectories from all animals and all gene segments. The flags `--dq_cut 0.90` (default 0.05) and `--dep_cut 100` were invoked, forcing SAMFIRE to retain trajectories changing by as much as 0.90 and to require a minimum read depth of 100 reads at each time point.

To infer the more conservative (i.e. larger) C-value, the SAMFIRE command `sl_neutrality` was employed to identify potentially non-neutral sites. A second C-value was then computed on the basis of neutral sites only.

Processing for Within-Host Selection Inference Preparing for within-host selection inference, data for the first two time points in the recipient animals (4295, 4297, 4298, 4302, 4306, 4309, 4310, 4314, 4315 and 4317) were collected and initial processing was repeated for these data. Next, the `sl_neutrality` command was called with the standard C-value identifying potentially non-

Table 5.5. Inferred within-host effective selection coefficients for the pig transmission dataset. Parameters were inferred for each segment in each individual recipient animal. Effective selection was estimated on the basis of the first two time points with coefficients reported to two decimal places.

Animal	Segment	Variant	Coefficient
4306	PB2	T230A	1.17
4306	PB2	T2129C	1.09
4309	HA	A716G	0.74
4314	MP	C384T	1.25
4314	NP	A1250G	1.19
4314	PB1	T1509A	−1.08
4317	HA	A511G	1.19
4317	HA	A619C	0.79

neutral trajectories. Multi-locus trajectories were then computed on the basis of the non-neutral trajectories using `call_ml` and a frequency cut-off of 2%.

Processing for Transmission Inference Preparing for transmission inference, data were collected matching the potential transmission events. Here, data from the last time point in the host were paired with data from the first time point in the recipient. The initial processing was then repeated for these data. Finally, multi-locus trajectories were called on the basis of the single-locus trajectories.

5.2.4.4 Inference of Effective Within-Host Selection

Estimation of within-host effective selection was carried out using inference code written by Chris Illingworth (Illingworth 2015). Preparing for this, within-host selection data were split by segment in order to infer effective selection at the level of individual genes. Given potentially non-neutral sites, an initial neutral inference was performed defining a likelihood baseline. Progressively, selection models of increasing complexity were constructed and evaluated, retaining models resulting in a BIC improvement of more than 10 log likelihood units. Within-host selection inference was terminated when no further improvements were found or if no additional models of increasing complexity could be generated. Inferences are shown in Table 5.5. Scripts for within-host selection inference can be found in the online repository, see Section 3.2.17.

5.2.4.5 Transmission Inference

Transmission inference was performed for the 28 potential transmission events under the assumption of neutral transmission and using a conservative C-value of $C = 186.33$. Inference methods accounting for and ignoring within-host selection were employed.

5.2.4.6 Phylogenetic Inference

Maximum likelihood phylogenetic trees were constructed for the transmission and within-host consensus sequences. Consensus sequences were obtained from SAMFIRE (Illingworth 2016) using the `consensus` command (SAMFIRE 1.05 and later). Phylogenetic inferences were obtained in MEGA7 (Kumar, Stecher and Tamura 2016) using the Hasegawa-Kishino-Yano model (Hasegawa, Kishino and Yano 1985).

5.3 Results

5.3.1 Transmission Inference

Employing an estimate of effective within-host selection, bottleneck inferences were computed using the neutral version of our model with outcomes shown in Figures 5.2 to 5.4. For comparison, bottleneck inferences in the absence of within-host selection were computed and displayed in Figures 5.2 to 5.4 only when the two outcomes differed. In general, the two approaches produced identical results, differing only in a small number of cases. In the following, if not specified otherwise, we refer to results obtained from accounting for within-host selection.

For the transmission from the four seeder pigs (4N) to the first set of recipients (6N-c1), narrow bottlenecks were generally inferred with medians falling in the range of 1–8 virions. Only two potential transmission events achieved bottlenecks of eight virions or more, namely 4301 $\rightarrow$ 4295 and 4305 $\rightarrow$ 4295, with eight virions being the theoretical limit identified in the Worby, Lipsitch and Hanage study, above which shared variant approaches outperform distance metrics. Under the hypothesis that the larger of the inferred bottleneck sizes indicates the most likely transmission link, the transmissions with the largest median bottlenecks were printed in bold. Given the narrow range of inferred bottlenecks it proved difficult to confidently identify true transmission events

with e.g. the sources of animals 4309 and 4317 being effectively unspecified. Regardless, the method provided some evidence for animal 4305 being the infector of 4298 and either 4301 or 4305 infecting 4295. Bottlenecks inferred in the absence of selection resulted in different outcomes only for transmissions 4303 $\rightarrow$ 4309 and 4301 $\rightarrow$ 4317 for which the median inferred bottleneck was marginally larger.

In the case of the potential transmission 4301 $\rightarrow$ 4317, an initial inference resulted in an infinite bottleneck prediction in 75% of the statistical replicates. Whilst biologically unrealistic, the infinite inferences are not a result of computational errors, but rather can arise as an artefact of the method itself; see Section 5.5 at the end of this chapter. Ignoring these inferences of infinity, a bottleneck of $N^T = 2$ was observed, whilst a re-analysis of transmission using $\Sigma^B = 0$ (see Section 5.5), resulted in a median bottleneck of $N^T = 1$.

Considering the four potential transmission events from the 6N-c1 host animals to recipient pig 4297 (3N-c1a), a more diverse range of bottlenecks were predicted as shown in Figure 5.3. For the transmission 4295 $\rightarrow$ 4297, a multimodal distribution of bottlenecks were inferred, having a median of $N^T = 57$. Manual inspection of the data revealed preservation of multiple variants across transmission, supporting the inference of a potentially large bottleneck size. For the remaining three transmissions, narrow distributions were inferred having means of $N^T = 14$ (4298 $\rightarrow$ 4297), $N^T = 5$ (4309 $\rightarrow$ 4297), and $N^T = 6$ (4317 $\rightarrow$ 4297). Taken together, this provides compelling evidence that the true infector of animal 4297 was in fact 4295.

Finally, narrow bottlenecks with medians of 1–7 virions were inferred for the 12 potential transmissions from the 4N seeder pigs to the 6N-c2 recipient animals. The largest median bottleneck size, $N^T = 7$, was inferred for transmissions 4301 $\rightarrow$ 4314 and 4303 $\rightarrow$ 4310. Infinite bottlenecks were predicted sporadically for the transmission 4303 $\rightarrow$ 4314. Ignoring inferences of infinity or reevaluating transmission with $\Sigma^B = 0$ (see Section 5.5) both resulted in median bottlenecks of $N^T = 1$. Some evidence was found for animal 4303 being the source of infection in pig 4310 and for 4301 being the infector of 4314. Median bottlenecks of 1–2 virions were predicted for transmissions to animals 4302 and 4306, making it impossible to identify the true transmission link in these cases. Ignoring effects of within-host selection resulted in different bottleneck inferences in three instances, namely transmissions 4305 $\rightarrow$ 4306, 4303 $\rightarrow$ 4314 and 4305 $\rightarrow$ 4314. Only the latter case resulted in a distinctly different outcome; here a multi-modal distribution with median $N^T = 12$ was predicted.

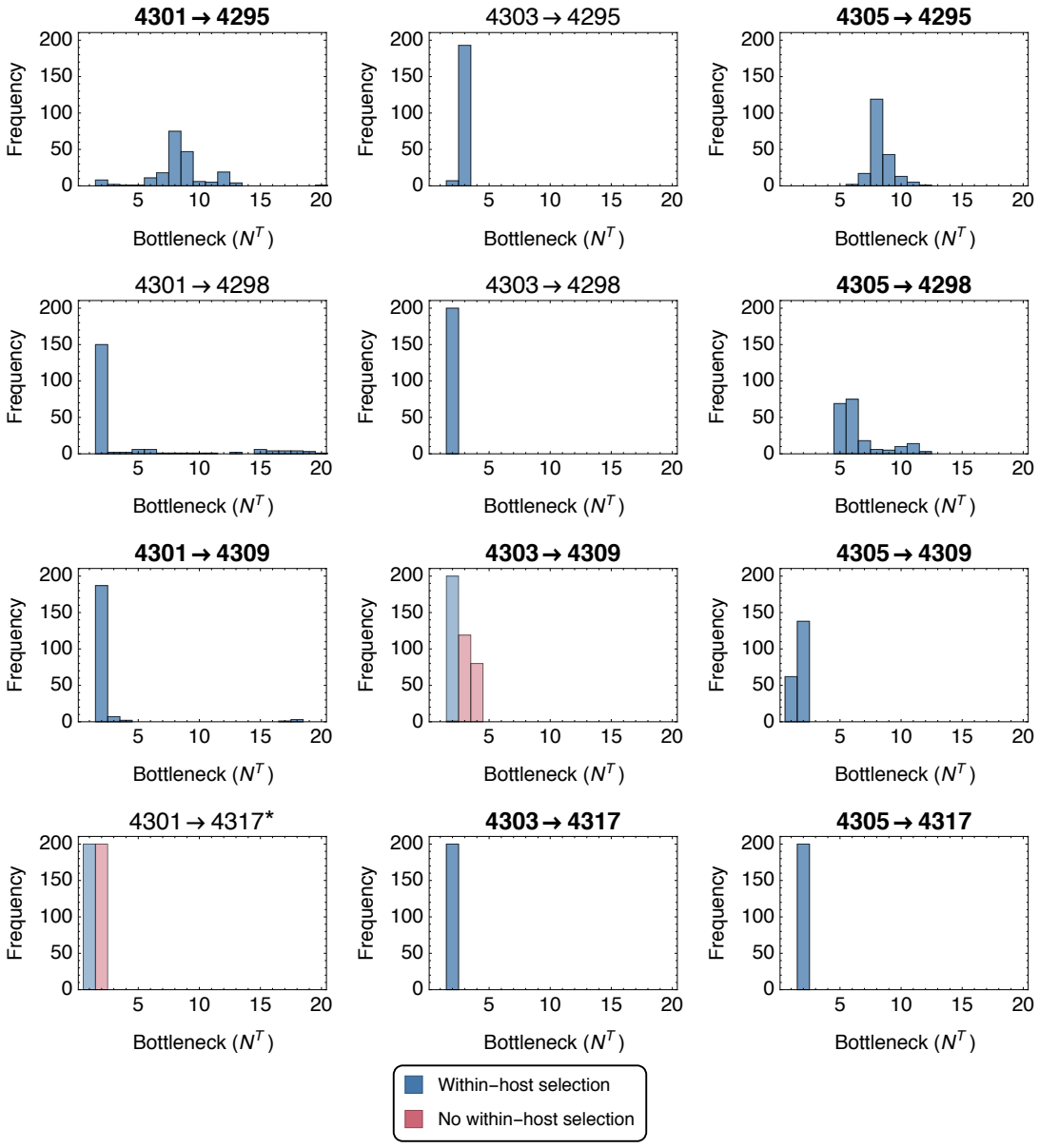


Figure 5.2. Histograms of bottleneck inferences for the 12 potential transmission events from the 4N seeder pigs to the 6N-c1 recipient pigs. Bottleneck inferences are based on 200 analysis seeds and inferences larger than $N^T = 20$ have been omitted for clarity. Inference methods either accounting for or ignoring within-host selection were employed; where these differ both outcomes have been shown. Bottleneck estimation for the transmission pair 4301 to 4317 (marked with an asterisk) resulted in inferences of infinity (see text for details) when accounting for within-host selection. Instead, analysis was repeated using a method not accounting for the variance in $\boldsymbol{q}^B$. Bold headings represent the most likely transmission events under the model accounting for selection.

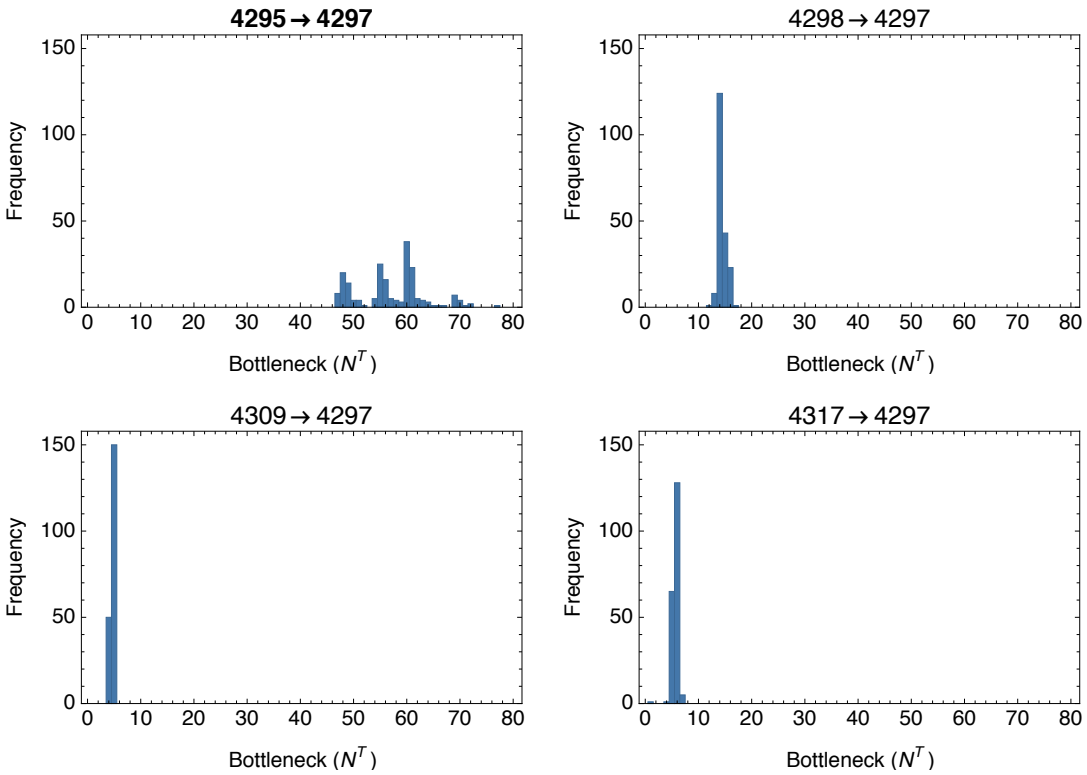


Figure 5.3. Histograms of bottleneck inferences for the four potential transmission events from the 6N-c1 seeder pigs to the 3N-c1a recipient pigs. Bottleneck inferences are based on 200 analysis seeds. Inference methods either accounting for or ignoring within-host selection were employed and resulted in identical outcomes. Bold headings represent the most likely transmission events.

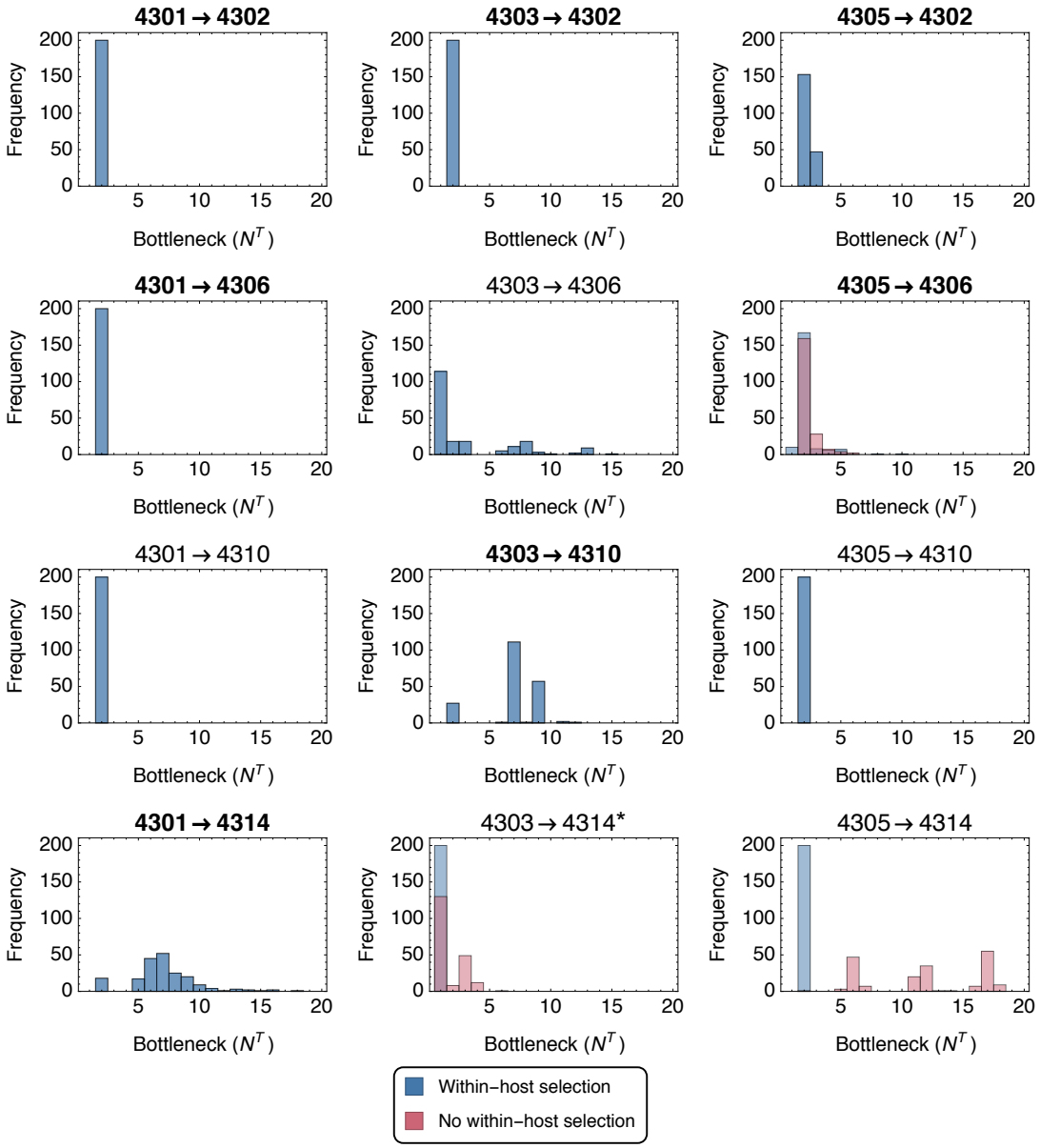


Figure 5.4. Histograms of bottleneck inferences for the 12 potential transmission events from the 4N seeder pigs to the 6N-c2 recipient pigs. Bottleneck inferences are based on 200 analysis seeds and inferences larger than $N^T = 20$ have been omitted for clarity. Inference methods either accounting for or ignoring within-host selection were employed; where these differ both outcomes have been shown. Bottleneck estimation for the transmission pair 4303 to 4314 (marked with an asterisk) resulted in inferences of infinity (see text for details) when accounting for within-host selection. Instead, analysis was repeated using a method not accounting for the variance in $\boldsymbol{q}^B$. Bold headings represent the most likely transmission events under the model accounting for selection.

Table 5.6. Predicted hosts for the nine recipient animals under the minimum genetics distance approach.

Recipient	Predicted host(s)
4295	4301, 4305
4297	4295, 4298, 4317
4298	4301, 4305
4302	4301, 4305
4306	4301, 4305
4309	4301, 4305
4310	4301, 4305
4314	4301, 4305
4317	4301, 4305

5.3.2 Phylogenetic Inference and Minimum Genetic Distance

Based on time points surrounding transmission, a maximum likelihood phylogenetic tree was constructed as shown in Figure 5.5. We observe a distinct lack of evolutionary signal, with most samples sharing a common consensus sequence, including seeder pigs 4301 and 4305. Of the 28 potential transmission events, 13 of them show no consensus sequence differences across transmission $\big(\{4301, 4305\} \rightarrow \{4295 \text{ dpi } 7, 4298 \text{ dpi } 7, 4309 \text{ dpi } 6, 4302, 4314\}$ and $\{4295 \text{ dpi } 8, 4298 \text{ dpi } 9, 4317 \text{ dpi } 7\} \rightarrow 4297\big)$. Consensus sequences for the remaining transmission events are only a few substitutions apart $\big(\{4301, 4305\} \rightarrow \{4306, 4310\}$ (one nucleotide difference), 4309 dpi 8 $\rightarrow$ 4297 (one difference), $4303 \rightarrow \{4295 \text{ dpi } 7, 4298 \text{ dpi } 7, 4309 \text{ dpi } 6, 4302, 4314\}$ (two differences), $\{4301, 4305\} \rightarrow 4317$ dpi 6 (three differences), $4303 \rightarrow 4310$ (three differences), $4303 \rightarrow 4306$ (four differences)), and $4303 \rightarrow 4317$ dpi 6 (five differences)$\big)$. Taken together, this suggests that a minimum genetic distance approach to inferring the underlying transmission network is likely to be inconclusive. For completeness, however, Table 5.6 shows the predicted transmission links obtained by minimising genetic distance between samples. This approach consistently predicts either 4301 or 4305 as the true infectors among the 4N seeder pigs. For the transmission to pig 4297 the method is more inconclusive, predicting either of 4295, 4298 and 4317 as the true source of infection.

The maximum likelihood phylogenetic tree for all available time points is shown in Figure 5.6. Exhibiting a flat hierarchy and with 20 out of 29 samples sharing the same consensus sequence, the full tree supports the absence of an

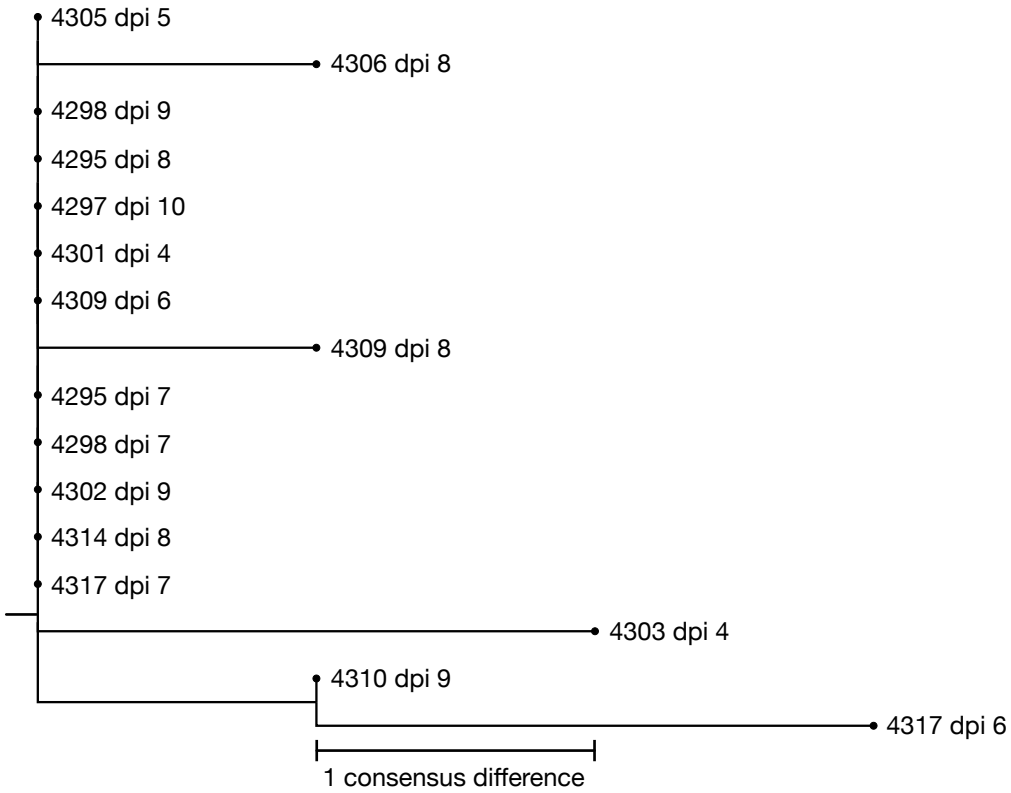


Figure 5.5. Maximum likelihood phylogenetic tree of consensus sequences for time points surrounding transmission events. The phylogeny was constructed in MEGA7 (Kumar, Stecher and Tamura 2016) using the Hasegawa-Kishino-Yano model (Hasegawa, Kishino and Yano 1985). The scale bar corresponds to a single consensus sequence difference. Some animals appear twice, e.g. 4317 for which both the recipient (6 dpi) and host (7 dpi) time points are shown.

Table 5.7. Absolute (Θ) and normalised (θ) sub-consensus sequence distance metrics for each of the 12 potential transmission events involving the four seeder pigs (4N) and the first set of recipient animals (6N-c1). The metrics corresponding to the most likely transmissions are shown in bold.

Recipient	4301		4303		4305	
	Θ	θ	Θ	θ	Θ	θ
4295	35.6	0.002 70	29.2	0.002 22	29.0	**0.002 20**
4298	31.8	0.002 41	23.1	0.001 75	21.9	**0.001 66**
4309	30.2	0.002 29	21.2	0.001 61	19.5	**0.001 48**
4317	31.8	0.002 41	22.0	0.001 67	21.1	**0.001 60**

evolutionary signal in the dataset. Generally, the within-host differences are as large as the between-host distances, suggesting that the transmission processes don't substantially alter the diversity, thus inhibiting a clear inference of route of transmission. Most striking, perhaps, is the within-host evolution of host animal 4303 and of 4317 (host and recipient animal). The consensus sequence of 4303 changes daily, starting one nucleotide difference away from the overall consensus on day 2, then changing to the overall consensus on day 3 and finally gaining two nucleotide differences on day 4. Similarly, for 4317 the consensus sequence reverts to the overall consensus at day 7, having been three nucleotide differences away on the previous day (first time point after transmission). Given the overall lack of an evolutionary signal, this behaviour is likely to be due to polymorphisms hovering around the 50% frequency mark.

5.3.3 Sub-Consensus Sequence Distance Metric

As an alternative to the minimum genetic distance approach, we computed absolute and normalised sub-consensus sequence distance metrics for the 28 potential transmission events, as shown in Tables 5.7 to 5.9. Whereas the consensus method predicted both 4301 and 4305 as the most likely hosts for transmissions to the 6N-c1 and 6N-c2 animals, the sub-consensus method favoured 4305 over 4301. Similarly, for transmission to 4297 the minimum genetic distance method considered 4295, 4298, and 4317 as being equally likely as host animals whilst the sub-consensus method favoured 4295. We note, however, that many of the sub-consensus metrics are of similar magnitudes, suggesting a degree of uncertainty in the conclusions presented here.

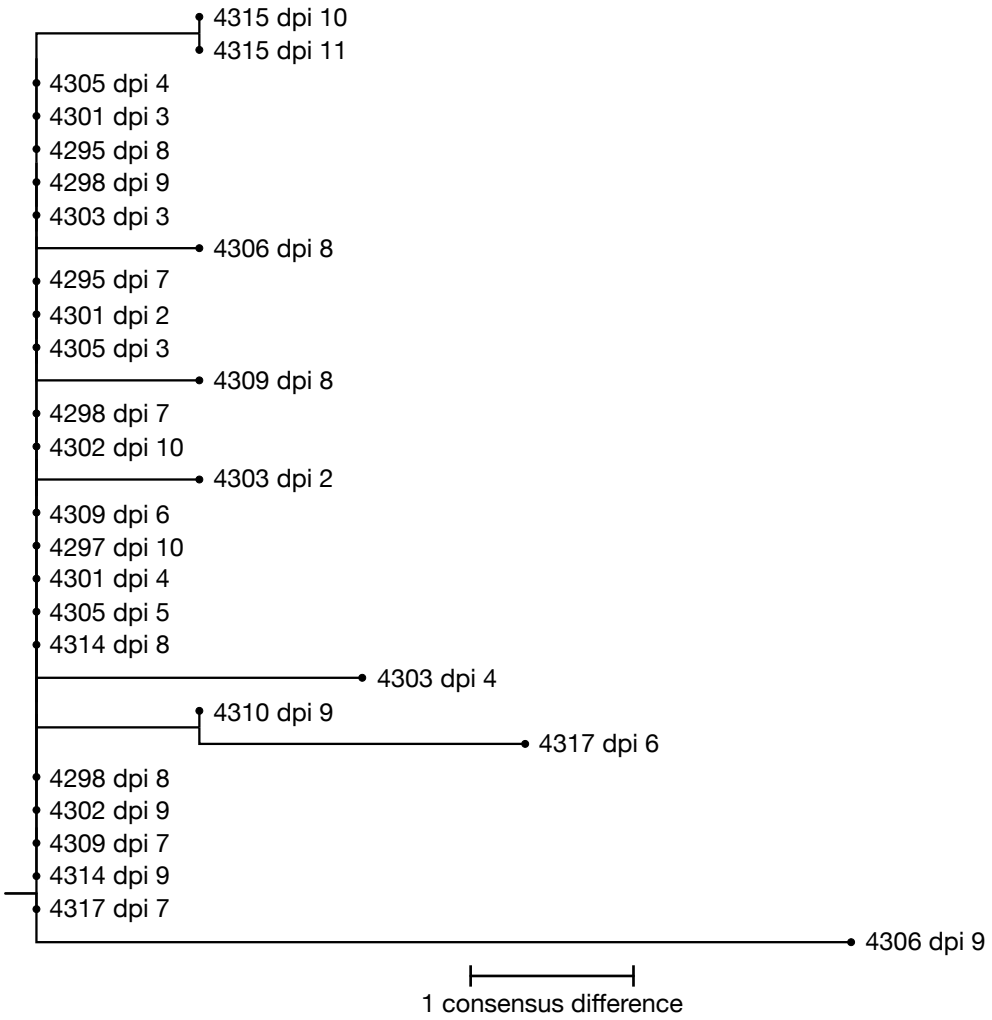


Figure 5.6. Maximum likelihood phylogenetic tree of consensus sequences for all within-host time points. The phylogeny was constructed in MEGA7 (Kumar, Stecher and Tamura 2016) using the Hasegawa-Kishino-Yano model (Hasegawa, Kishino and Yano 1985). The scale bar corresponds to a single consensus sequence difference. Time points with low coverage (4297 dpi 11, 4305 dpi 2, 4315 dpi 9) were ignored.

Table 5.8. Absolute (Θ) and normalised (θ) sub-consensus sequence distance metrics for each of the four potential transmission events involving the first set of recipient animals (6N-c1) and the secondary set of contact animals (3N-c1a). The metric corresponding to the most likely transmission is shown in bold.

Recipient	4295		4298		4309		4317	
	Θ	θ	Θ	θ	Θ	θ	Θ	θ
4297	30.3	**0.002 30**	32.3	0.002 45	32.2	0.002 44	30.6	0.002 32

Table 5.9. Absolute (Θ) and normalised (θ) sub-consensus sequence distance metrics for each of the 12 potential transmission events involving the four seeder pigs (4N) and the second set of recipient animals (6N-c2). The metrics corresponding to the most likely transmissions are shown in bold.

Recipient	4301		4303		4305	
	Θ	θ	Θ	θ	Θ	θ
4302	32.6	0.002 47	23.0	0.001 75	22.5	**0.001 71**
4306	30.3	0.002 29	20.4	0.001 55	18.9	**0.001 43**
4310	31.8	0.002 41	22.5	0.001 71	21.3	**0.001 61**
4314	32.1	0.002 44	23.8	0.001 81	21.5	**0.001 63**

Table 5.10. Absolute (Θ) and normalised (θ) sub-consensus sequence distance metrics based on SAMFIRE filtered variants for each of the 12 potential transmission events involving the four seeder pigs (4N) and the first set of recipient animals (6N-c1). The metrics corresponding to the most likely transmissions are shown in bold.

Recipient	4301		4303		4305	
	Θ	θ	Θ	θ	Θ	θ
4295	9.11	0.0939	6.25	0.120	5.94	**0.0849**
4298	8.04	0.0759	5.61	0.0891	5.02	**0.0612**
4309	7.05	0.0850	4.04	0.106	3.40	**0.0630**
4317	9.08	0.111	5.72	0.147	5.49	**0.0998**

Aiming to filter out potential noisy observations, we computed sub-consensus sequence distance metrics on the basis of a filtered set of single-locus variants identified by SAMFIRE. As shown in Tables 5.10 to 5.12, variants identified by SAMFIRE as polymorphisms correspond on average to just over 20% of the absolute metric across all sites. More importantly, though, the normalised metrics for the SAMFIRE variants are on average almost 50 times larger than the equivalent metric across all sites. This suggests that the identified polymorphisms are potentially more informative with regards to transmission and that the signal from these may be masked by noise. Additionally, it should be noted that the unfiltered metrics may be biased as sequencing artefacts located in regions of low read coverage may have a disproportionate amount of impact on the metric. Based on the SAMFIRE variant sub-consensus metrics, identical conclusions were drawn, but with greater confidence in the inferences; compare for instance Tables 5.8 and 5.11. These inferences agree to a large extent with the outcomes predicted by the bottleneck route of transmission approach.

Table 5.11. Absolute (Θ) and normalised (θ) sub-consensus sequence distance metrics based on SAMFIRE filtered variants for each of the four potential transmission events involving the first set of recipient animals (6N-c1) and the secondary set of contact animals (3N-c1a). The metric corresponding to the most likely transmission is shown in bold.

Recipient	4295		4298		4309		4317	
	Θ	θ	Θ	θ	Θ	θ	Θ	θ
4297	2.28	**0.0486**	2.41	0.0548	4.25	0.0817	3.37	0.0716

Table 5.12. Absolute (Θ) and normalised (θ) sub-consensus sequence distance metrics based on SAMFIRE filtered variants for each of the 12 potential transmission events involving the four seeder pigs (4N) and the second set of recipient animals (6N-c2). The metrics corresponding to the most likely transmissions are shown in bold.

Recipient	4301		4303		4305	
	Θ	θ	Θ	θ	Θ	θ
4302	8.85	0.0972	5.88	0.128	5.24	**0.0845**
4306	7.69	0.0884	4.12	0.106	3.91	**0.0674**
4310	8.49	0.0975	5.27	0.123	5.03	**0.0824**
4314	7.43	0.0808	5.12	0.0985	3.80	**0.0594**

Table 5.13. Number of shared variants for each of the 12 potential transmission events involving the four seeder pigs (4N) and the first set of recipient animals (6N-c1). The most likely transmissions are shown in bold.

Recipient	4301	4303	4305
4295	3	**7**	3
4298	**4**	3	1
4309	5	**6**	5
4317	**6**	5	4

Table 5.14. Number of shared variants for each of the four potential transmission events involving the first set of recipient animals (6N-c1) and the secondary set of contact animals (3N-c1a). The most likely transmissions are shown in bold.

Recipient	4295	4298	4309	4317
4297	**4**	**4**	2	**4**

5.3.4 Route of Transmission From Shared Variants

As a final attempt at determining route of transmission we considered the maximum shared variant approach of Worby et al. (2017). For our purpose, shared variants were defined as across-transmission single-locus polymorphisms for which minor alleles at both time points accounted for a frequency of minimum 2% and were supported by at least 10 reads. Inferring here the most likely transmission links by maximising the number of variants shared between host and recipient, we obtained the results shown in Tables 5.13 to 5.15. These results differ somewhat from those of both the minimum genetic distance approach and the sub-consensus sequence metric. Considering for instance recipient 4295 (Table 5.13), the shared variant method has a strong preference for host 4303, whilst the sub-consensus method is in favour of animal 4305 and the minimum distance method prefers pigs 4301 and 4305. The remaining transmission links in Table 5.13 are less strongly predicted, with the preferred animal pair sharing only one variant more than the alternatives.

A similarly non-specific picture was found when considering recipient 4297 (Table 5.14) for which 4295, 4298, and 4317 were equally likely infectors under the shared variant approach. This agrees well with the findings of the minimum genetic distance approach, which favoured the same set of host animals. Conversely, the sub-consensus and bottleneck approach both identified a clear signal for 4295 being the true infector.

Table 5.15. Number of shared variants for each of the 12 potential transmission events involving the four seeder pigs (4N) and the second set of recipient animals (6N-c2). The most likely transmissions are shown in bold.

Recipient	4301	4303	4305
4302	7	**9**	7
4306	6	**10**	6
4310	8	**9**	5
4314	**8**	6	**8**

Finally, considering the potential transmissions from the four seeder pigs to the second set of recipient animals (Table 5.15), the maximum shared variant method strongly favoured animal 4303 as the infector of 4306. This result was not reflected by the sub-consensus method, which preferred 4305 as the source of infection, and the minimum distance method, which attributed transmission due to 4301 and 4305. For the remaining recipient animals the maximum shared method was less convincing, with the preferred host-recipient pairs sharing one or two more variants than the alternatives.

Overall a mixed picture emerges, suggesting either that one of the route of transmission methods is significantly more accurate than the others, or, perhaps more likely, that the transmission signal in most cases is too weak for reliably estimating transmission networks.

5.4 Discussion

In this chapter I introduced the concept of effective selection and employed it for the inference of bottlenecks in a transmission study in pigs. Within-host effective selection was estimated by fitting a multi-locus model of evolution to the first two time points in each of the recipient animals, thus capturing the essence of selection by assuming the underlying fitness landscape remaining approximately constant during short time intervals. Whilst present in only a subset of the individuals, incorporation of effective selection allowed for a potentially unbiased inference of bottleneck size. In general, narrow bottlenecks in the range of 1–8 virions were inferred, being in agreement with our predictions for the ferret dataset (Moncla et al. 2016) and current estimates for influenza transmission in humans (McCrone et al. 2018). Two inferences resulted in substantially larger bottleneck sizes, namely the transmissions from host animals 4295 and 4298 to recipient pig 4297 for which median bottlenecks of 57 and 14 viruses were

predicted respectively. Manual inspection of data supported the validity of these inferences. Whilst generally narrow, this suggests that influenza transmission bottlenecks in mammals may occasionally be considerably larger, perhaps owing to a close contact transmission setup. To our knowledge, this is the first time viral bottlenecks have been quantified from transmission events in pigs.

Taking a further look at effective selection, I performed bottleneck inferences both in the presence and absence of effective within-host selection. This resulted in very similar inferences with the two approaches differing only in a handful of instances. Where different, the bottleneck inferences accounting for effective selection tended to predict a marginally narrower bottleneck size. Being unable to account for selection for increased transmissibility, it is possible that the two measures of selection balance each other out to a certain degree. In general, we note that the use of effective selection warrants further investigation. As the method attempts to simultaneously account for inherent selection and selection due to interference, the method relies on the approximate constancy of linkage disequilibrium during short time intervals. In the event that the amount of linkage disequilibrium between alleles at the first two time points in the recipient differs substantially from the linkage disequilibrium at the founder time point, the estimate of effective selection may be incorrect. Being unable to properly account for the underlying multi-segment haplotype structure makes it difficult to investigate the validity of effective selection measures. Currently, I believe that the framework outlined here represents the best possible solution to an otherwise impossible problem.

I posited the idea of route of transmission inference based on differences in bottleneck size estimations. In particular, where multiple potential transmission events exist, the transmission resulting in the largest inferred bottleneck size will in general represent the most likely true transmission link. Within this framework, strong evidence was found for animal 4295 being the true infector of pig 4297. Additionally, the method provided moderate evidence for transmissions between animals 4305 and 4298, 4303 and 4310, 4301 and 4314, and either 4301 or 4305 infecting 4295. About half of the transmission links were undefined under this method, owing to the inference of narrow bottlenecks for all potential transmission events. Whilst this approach warrants further investigation, the ability of the method to simultaneously account for noise, selection and within-host evolution, as well as being inherently rooted in a sub-consensus framework, suggests the potential for the method to perform well in cases where transmission inference results in differential bottleneck predictions.

Seeking potential alternative explanations of the data, I explored route of transmission inference from previously established methods. Inferring a phylogeny for the transmission dataset indicated a high degree of similarity at the consensus sequence level, suggesting that minimum genetic distance approaches may be inconclusive. Indeed, minimising genetic distance at the consensus level resulted in identical inferences of host animals for all potential transmission links, with inferences favouring two or more hosts. This highlights the inability of consensus level methods to correctly capture route of transmission characteristics from outbreaks of acute infectious disease.

Considering instead the shared variant approach of Worby et al. (2017), which represents a sub-consensus approach to route of transmission inference, a more diverse picture emerged. Here, considerable evidence was found for animal 4303 infecting 4295 and for pig 4303 being the true infector of 4306. Whilst additional transmission links were deduced, these were less well supported due to a limited number of differences in the number of shared variants. Interestingly, the transmission links predicted here are in general irreconcilable with the conclusions drawn from the bottleneck approach. For instance, for recipient animal 4295 the bottleneck method favoured host animals 4301 and 4305 whilst the shared variant method preferred animal 4303. Assuming the bottleneck inferences themselves ($N^T = \{8, 3, 8\}$ for 4301, 4303 and 4305 respectively) are valid, the inconsistency may be understood by noting that the shared variant method underperforms in cases of narrow bottlenecks. Additionally, the bottleneck inference approach considers a significantly larger set of single locus variants (57-105 variant sites) than the shared variant method (3-7 variants). This is expected in the case of narrow bottlenecks where most polymorphisms go extinct across transmission. As such, the bottleneck approach is increasingly sensitive, considering vastly more data than the shared variant method. Regardless, the two methods may potentially be used in concert; where large bottlenecks of similar magnitudes are inferred using the bottleneck approach, the shared variant method might be employed to further differentiate the transmission links.

As a final approach to route of transmission inference we proposed a sub-consensus sequence distance metric, aiming on the one hand to account for sub-consensus signatures and, on the other, to provide a computationally efficient alternative to the bottleneck inference method. Considering both genome-wide data and a smaller set of filtered variants, the sub-consensus metric gave identical conclusions, albeit with the filtered variants providing additional support for the specific findings. Interestingly, the method agreed well with the bottleneck

inference approach, differing only in the identification of transmission sources for animals 4310 and 4314. The latter inconsistency may be explained by the presence of within-host selection for animal 4314; by neglecting selection, the sub-consensus sequence distance metric may incorrectly identify 4305 as the most likely infector of 4314. Indeed, when bottleneck inference was carried out without accounting for within-host selection, this method also predicted 4305 to be the true infector of 4314. Regardless, strong evidence was found for 4295 being the infective source of 4297, suggesting that the two methods are aligned where the signal is strong. In conclusion, the sub-consensus sequence metric provides a simple alternative to performing an in-depth transmission analysis when simply aiming to elucidate transmission networks. A subsequent detailed analysis of transmission may still be required in order to correctly account for the many factors potentially impacting viral transmission.

5.5 Appendix on Infinite Bottleneck Inferences

Occasionally the transmission inference method may predict infinitely large bottlenecks. This may occur if the observations of the system before and after transmission are very similar; as observed previously, if the extent of noise in the observations, coded by C, is overestimated, the observed difference between samples collected before and after transmission may be smaller than the expected difference between samples arising by noise; in such a case an arbitrarily high inferred bottleneck is the result.

In the calculations of this chapter, we sometimes identified cases where the bottlenecks inferred under different replicate calculations, with different reconstructions of the pre-transmission population, gave either a low bottleneck of order around 10, or an infinitely high bottleneck. We here show that a high bottleneck may be inferred from a 'bad' reconstruction of the pre-transmission population. Considering a one-dimensional system with pre-transmission mean μ^B and standard deviation σ^B, the compound distribution for the founder population under a neutral transmission becomes (cf. Equation C.4)

$$\text{var}[q^F] = \frac{1}{N^T}\mu^B(1-\mu^B) + \left(1-\frac{1}{N^T}\right)\left(\sigma^B\right)^2 \tag{5.4}$$

The variance in q^F is made up of two components; one accounting for the uncertainty arising from the binomial transmission step, $\frac{1}{N^T}\mu^B(1-\mu^B)$, and one due to the variance inherent in q^B, $\left(1-\frac{1}{N^T}\right)\left(\sigma^B\right)^2$. The bottleneck N^T defines

the balance between these two components. When the bottleneck is small, the variance is primarily due to the bottleneck itself and the first term is dominant, i.e. $\frac{1}{N^T} > \left(1 - \frac{1}{N^T}\right)$ for small N^T. Conversely, when the bottleneck is large, the compound variance in q^F is defined by σ^B, i.e. the precision with which μ^B is specified, rather than the binomial variance.

In a one-dimensional system we generally expect the compound variance to decrease as N^T is increased. From inspection, this is the case only when $\mu^B(1-\mu^B) > \left(\sigma^B\right)^2$. For a single dimension this criteria can be controlled for during the optimisation of μ^B and σ^B, however, in a multi-dimensional setting a similar criteria cannot be established, lacking the ability to compare magnitudes of matrices. Regardless, it can easily be appreciated that certain sub-dimensional projections result in a similar criteria, where, as a result, the compound variance of q^F increase with N^T for specific optimisations of μ^B and σ^B.

Infinite bottleneck inferences may therefore, in some cases, be an entirely mathematical artefact arising from unrealistic optimisations of $\boldsymbol{\mu}^B$ and Σ^B. As a consequence, when infinite bottleneck inferences were encountered, I reevaluated the transmission event using a pre-transmission population with $\Sigma^B = 0$. I have previously shown that this approach is valid, albeit leading to slight underestimations of bottleneck sizes for large N^T, see Section 3.3.2.

Chapter 6

Analysis of Transmission Data From Human Challenge Study

6.1 Introduction

In the previous chapter I demonstrated the ability of my inference method to estimate population bottlenecks from transmission events in small herds of swine. Accounting for potentially modest levels of viral reassortment, inferences considered differing within-host haplotype fitnesses by appealing to a framework of effective selection. In this chapter I consider transmission inference from influenza infections arising from human challenge studies, presenting here a mathematically elegant approach to within-host selection rooted in an existing multi-segment, multi-locus inference of a fitness landscape. Based on this, I demonstrate the ability of my method to infer important parameters of transmission from human infections, discussing potential limitations and challenges associated with transmission inference from experimental data.

6.1.1 Human Challenge Studies

Human challenge studies (also known as controlled human infection model studies, deliberate exposure, etc.) have obtained an increased focus in recent years. Human challenge studies refer to the purposeful exposure of human volunteers to infectious agents in a controlled and carefully monitored environment (Darton et al. 2015). Challenge studies provide a range of scientific and clinical advantages that the alternatives, such as animal models, simply do not possess; while animal models can replicate many aspects of human infection, they may

fall short of the reality in other ways. A key use of human challenge studies is for the identification of vaccine candidates and estimation of vaccine efficacy (Darton et al. 2015; McArthur and Shirley 2011). Surpassing traditional clinical trials, challenge studies allow vaccines to be developed at a rapid pace, and enable the identification of unviable vaccine candidates at an early stage, saving time, money and resources. Human challenge studies are particularly useful for host-restricted pathogens, for which an appropriate animal model may not exist, and in cases where infection in animal models does not properly emulate disease progression in humans. One example of this is dengue virus, where the infection of non-human primates result in asymptomatic infections (Thomas 2013).

Human challenge studies allow disease development to be studied under highly controlled conditions, by fixing environmental surroundings such as temperature and humidity, the degree of interaction with other study participants, and restricting or allowing contact with the outside world. They may be of relevance where natural disease outbreaks occur either infrequently or sporadically, or where study of a disease in its natural host is limited by pathogen seasonality (Hirve et al. 2016). Taking a more fundamental view, challenge studies may be employed for the verification of cause-and-effect relationships between a pathogen and a clinical disease, providing the most direct approach for demonstrating Koch's postulates of disease causation (Darton et al. 2015; Koch 1876, 1984).

In recent years, human challenge studies have aided the development of vaccines, for example by reducing the need for time-consuming phase III trials (Darton et al. 2015; McArthur and Shirley 2011). One 2017 human challenge study considered the efficacy of a typhoid conjugate vaccine as an alternative to currently licensed typhoid vaccines. Powered by 103 adult study individuals, the conjugate vaccine was found to have an efficacy of 87%, which was considerably larger than the 52% efficacy obtained for the comparator, an established Vi-polysaccharide vaccine (Jin et al. 2017). This study prompted the World Health Organization to update its position on typhoid preventives, now identifying the conjugate vaccine as the preferred vaccine for individuals of all ages (World Health Organization 2018c). With a view to vaccine development, the Wellcome Trust has recently called for an expansion of the number of human challenge studies (The Wellcome Trust 2018a,b).

Human challenge studies are by no means a new concept, with challenge studies dating as far back as the 18th century (Darton et al. 2015). A key milestone, of course, was the invention of vaccines, with which Edward Jenner is often credited based on his work on smallpox in the late 1700s. By directly inoculating a

young boy, James Philips, with cowpox virus, Jenner proved that exposure to cowpox generates protection against the significantly more dangerous smallpox virus (Riedel 2005). Historically, human challenge studies have also been carried out in the context of influenza (Carrat et al. 2008), of which the most striking experiments involved direct inoculation of volunteers with virus deriving from the 1918 pandemic (Wahl, White and Lyall 1919; Yamanouchi, Sakakami and Iwashima 1919).

As illustrated by the above examples, human challenge studies raise important ethical questions regarding patient safety; at its core, challenge studies are at odds with fundamental Hippocratic principles of non-maleficence. As a consequence, modern human challenge studies have strictly defined ethical and safety guidelines, with the safety of volunteers, staff and the general population being the primary concern (Bambery et al. 2016; The Academy of Medical Sciences 2005). For example, challenge studies aren't to be conducted in diseases for which no effective treatment exists (Bambery et al. 2016).

Finally, whilst generally versatile, a handful of limitations must be considered in relation to human challenge studies. Firstly, challenge studies often takes place in developed countries wherein the challenge population consists of healthy, adult volunteers, who generally haven't been exposed to the disease prior to the experiment. This may result in a substantial gap between the study and target populations when considering diseases that primarily infect individuals in developing countries. For instance, in the case of cholera, where the target population includes children living in unsanitary, impoverished regions where the disease is endemic, vaccine efficacy may be overestimated (McArthur and Shirley 2011). Secondly, the experimentally induced infection may differ substantially from the naturally occurring infection. Often, to ensure successful infection, a large inoculation dose may be applied, sometimes consisting of strains with an increased severity compared to the wild type. Studies in ferrets and humans suggest that influenza infection initiated by intranasal inoculation may not sufficiently represent natural infection, e.g. with intranasal inoculation resulting in decreased viral shedding, limited aerosol transmission and differing degrees of infection of the lower respiratory tract (Carrat et al. 2008; Gustin et al. 2011). While in some ways ferrets provide a valuable model of human influenza infection (Buhnerkempe et al. 2015), the study upon which this chapter builds highlighted a difference in the rate of reassortment between human and small animal infections arising from the apparent local density of infectious particles at the onset of host infection. In this chapter I investigate data from a human challenge

study in influenza and consider the inherent degree of realism achieved by these studies when compared to infections arising from natural viral transmission.

6.1.2 Author Contributions

The work presented in this chapter is currently unpublished. The work described here was carried out by the author under the supervision of his PhD supervisor, Dr Christopher Illingworth.

6.2 Methods

6.2.1 Weighted Within-Host Selection

Given a limited number of polymorphic loci, methods exist for the inference of within-host selection during influenza infection, accounting for between-segment correlations by jointly fitting selection and viral reassortment rates (Sobel Leonard et al. 2017a); such a calculation indicated that the effective within-host rate of reassortment is low in the individuals studied. This poses a challenge for our transmission model; in the previous chapters the assumption of rapid reassortment facilitated our approach to modelling within-host selection. To model within-host selection in this case we devised the principle of *weighted selection*. Similarly to the previous chapter we derived an effective within-host fitness for each segment within each host. However, in this case our effective fitness values were derived from a known within-host fitness landscape and reconstructed whole genome frequencies using a weighted mean approach.

From the previous analysis of the data, described in Sobel Leonard et al., we have an inferred multi-segment fitness landscape $\boldsymbol{\sigma}^{\mathrm{MS}}$, in which the parameter σ_i^{MS} describes the within-host fitness of the multi-segment haplotype i. Furthermore, for each host j we have the inferred within-host multi-segment haplotype frequencies $\boldsymbol{q}^{\mathrm{MS},j}(t)$, describing the evolution of each within-host population in terms of multi-segment haplotypes, under the influence of selection, mutation, and reassortment.

From these previously derived statistics, we wish to calculate the effective within-host fitness of each single-segment haplotype within each host. This statistic, being a function of the inherent fitness of the haplotype and linkage

disequilibrium with selected variants on other segments, will change over time as the population evolves. To consider within-host growth immediately following transmission we wish to evaluate this statistic at the earliest time point for which we have data.

In the individual j we calculate the weighted selection for a single-gene haplotype h_i as

$$\sigma_{i,j}(t) = \frac{\sum_{k:h_i \subset h_k^{\mathrm{MS},j}} q_k^{\mathrm{MS},j}(t)\sigma_k^{\mathrm{MS}}}{\sum_{k:h_i \subset h_k^{\mathrm{MS},j}} q_k^{\mathrm{MS},j}(t)} \tag{6.1}$$

where the sum is over all multi-segment haplotypes k which have the same alleles as the single-segment haplotype h_i. For example, if h_i described the alleles AT in haemagglutinin, the sum would be conducted over all multi-segment haplotypes with these alleles in haemagglutinin, irrespective of which alleles they contained in other segments. Noting the time-dependence of the statistic, the calculation of $\sigma_{i,j}(t)$ is performed in each case at the first time t for which a reconstruction of the viral population in individual j has been calculated; this is equal to the first time following transmission at which sequence data describing the viral population was collected.

6.2.2 Human Challenge Study

The analysis performed by Sobel Leonard et al. was conducted on data from two experiments, described respectively by Zaas et al. (2009) (experiment 1) and McClain et al. (2016) (experiment 2). Data from the former experiment, which extend far beyond those collected by viral sequencing, have been studied extensively, primarily for the classification of respiratory viral infections by host signatures based on microarray or RT-PCR gene expression profiling (Huang et al. 2011; Woods et al. 2013; Zaas et al. 2009, 2013), but also for the comparison of experimental infection and infection induced by vaccination (Moody et al. 2011) and for relating T cell responses to disease severity (Wilkinson et al. 2012). An earlier analysis of the data used changes in sequence diversity to identify the presence of selection acting at some point in the transmission process in experiment 1 (Sobel Leonard et al. 2016). This earlier study did not attempt to analyse the population bottleneck active in the transmission process.

6.2.2.1 Outline of Study

The first experiment, conducted in 2009, involved the challenge of 17 immunologically naive individuals with a H3N2 influenza of strain A/Wisconsin/67/2005. The viral stock was created from a detailed passaging pipeline, with the virus first passaged in avian primary chicken kidney cells, then in embryonated chicken eggs, and, finally, in GMPVero cells (Sobel Leonard et al. 2016). The viral stock was sampled prior to human challenge; we refer to the pre-challenge time point as $t = 0$. Volunteers were then intranasally challenged with the stock virus, dosages varying between $10^{3.08}$ and $10^{6.41}$ $TCID_{50}$ (Moody et al. 2011). Nasal washes were collected from individuals on a daily basis and oral oseltamivir was administered on day 6 (Zaas et al. 2009). Of the 17 patients in the study, seven of these had at least one successfully sequenced sample of which four individuals (P1001, P1006, P1012, and P1013) accounted for multiple time points. The samples for which sequence data are available are shown in Table 6.1. The viral stock contained 45 variants with frequencies above 2% (HA: 6, MP: 11, NA: 1, NP: 2, NS: 2, PA: 13, PB1: 7, PB2: 3), the variants being well distributed across the genome (Table 6.2).

In the second experiment, conducted in 2016, 21 immunologically naive volunteers were intranasally inoculated with influenza A virus, the viral stock being identical to that of the first experiment. The inoculation dose was fixed at 10^6 $TCID_{50}$ and nasal lavage samples were obtained on a daily basis. In order to compare treatment procedures, oral oseltamivir was administered either at 36 hours after inoculation or at 5 days post inoculation (McClain et al. 2016). Sequencing of nasal wash samples resulted in sequence data from ten of the 21 patients of which nine were sampled at multiple time points (Sobel Leonard et al. 2017a). An overview of the available samples for experiment 2 can be found in Table 6.3.

6.2.2.2 Calculation of Weighted Within-Host Selection

In a previous examination of the human challenge data, an advanced model of within-host evolution was employed for the estimation of a multi-segment characterisation of within-host fitness (Sobel Leonard et al. 2017a, Table 3 and Fig 3). The fitness landscape was produced from sequence data processed by SAMFIRE (Illingworth 2016) using a frequency cut-off of 2%. Additionally, multi-segment haplotypes and associated frequencies were estimated across the available time points. Using the frequencies at the first available time point, we

Table 6.1. Sampling times for individuals in experiment 1 of human challenge study. Times are with respect to inoculation (time=0). Individuals are denoted by their four digit patient ID, e.g. P1001.

Time	1001	1006	1008	1010	1012	1013
1	×	×	.	.	.	.
2	×	.	×	.	×	×
3	.	×	.	×	×	×
4	.	.	.	.	.	.
5	.	.	.	.	.	.
6	.	.	.	.	×	.

Table 6.2. Inoculum variant sites with frequencies in excess of 2%.

	HA	MP	NA	NP	NS	PA	PB1	PB2
Variant sites	530	10	826	357	337	85	1383	659
	641	12	.	898	552	88	1614	704
	1217	31	.	.	.	100	2265	706
	1272	55	.	.	.	130	2279	.
	1360	272	.	.	.	139	2282	.
	1584	725	.	.	.	145	2284	.
	.	945	.	.	.	1549	2291	.
	.	1016	.	.	.	1690	.	.
	.	1018	.	.	.	2002	.	.
	.	1020	.	.	.	2035	.	.
	.	1021	.	.	.	2038	.	.
	.	.	.	.	.	2116	.	.
	.	.	.	.	.	2122	.	.

Table 6.3. Sampling times for individuals in experiment 2 of human challenge study. Times are with respect to inoculation (time=0). Individuals are denoted by their four digit patient ID, e.g. P5001.

Time	5001	5002	5004	5006	5007	5017	5018	5019	5020	5021
1	.	.	.	×	×	.	×	×	×	×
2	×	×	×	×	.	×	×	×	×	.
3	×	×	×	×	.	.	.	×	×	×
4	.	.	×	.	.	.	.	.	×	×
5	×	.	×	.	.	.	.	.	.	×

here computed weighted within-host selection coefficients for each of the human challenge subjects by appealing to the framework outlined in Section 6.2.1. Inferences are shown in Tables 6.4 to 6.6. We note that the selective pressures are generally negative (with respect to the variant, i.e. non-consensus, haplotypes). This is in agreement with Sobel Leonard et al. (2017a) who found selection to be of a purifying nature; a number of polymorphisms are allowed to accumulate during initial passaging, but are subsequently found to be unfit when the virus enters the human host.

6.2.2.3 Data Processing

Pre-processed data from the previous study (Sobel Leonard et al. 2017a) were used for this analysis. As described in the paper, filtering of aligned reads was conducted using the SAMFIRE package, which was further used to infer the extent of noise in the data for the within-host analysis.

For the purposes of this study single-locus variants were identified using a frequency cut-off of 2%. An appropriate value of noise for the transmission analysis was derived as in previous chapters ($C = 138.86$; value for within-host analysis $C = 44.28$). To avoid a biased inference, duplicate trajectories (defined as trajectories covering the same loci, i.e. multiallelic loci) were removed, retaining the trajectory with the largest average polymorphism. The noise inference was then computed using the SAMFIRE command `sl_noise` on the basis of trajectories from all individuals and all gene segments. The flags `--dq_cut 0.90` (default 0.05) and `--dep_cut 100` were invoked, forcing SAMFIRE to retain trajectories changing by as much as 0.90 and to require a minimum read depth of 100 reads at each time point.

To infer the more conservative (i.e. larger) C-value, the SAMFIRE command `sl_neutrality` was employed to identify potentially non-neutral sites. A second C-value was then computed on the basis of neutral sites only.

Processing for Transmission Inference Preparing for transmission inference, data from the inoculum sample were paired with data from the first available time point in each of the study subjects. The initial processing was then repeated for these data. Finally, multi-locus trajectories were called on the basis of the single-locus trajectories.

Table 6.4. Weighted within-host selection coefficients for patients P1001-P1013. Selection coefficients (σ) are with respect to (potentially gapped) single-segment haplotypes defined by the variant loci in HA, NP, and PA. For example, any viral haplotype in patient P1001 exhibiting an A at position 545 and a G at position 1287 is under selective pressure by $\sigma = -1.62$.

Patient ID	HA				NP				PA	
	545	920	1287	σ	372	762	913	σ	1704	σ
P1001										
	A	-	G	−1.62	T	G	-	−4.05	C	−7.78
	A	-	A	−8.78	T	A	-	−7.18	T	−2.13
	C	-	G	−1.05	C	G	-	−6.09	.	.
	C	-	A	−9.62	C	A	-	−22.13	.	.
P1006										
	A	-	G	−1.87	T	G	G	−10.64	C	−12.61
	A	-	A	−12.24	T	G	A	−2.21	T	−2.13
	C	-	G	−1.06	T	A	G	−0.87	.	.
	C	-	A	−20.35	T	A	A	−6.59	.	.
	.	.	.	.	C	G	G	−3.01	.	.
	.	.	.	.	C	G	A	−6.36	.	.
	.	.	.	.	C	A	G	−0.96	.	.
	.	.	.	.	C	A	A	−6.36	.	.
P1012										
	A	C	G	−2.16	T	A	G	−0.87	C	−16.37
	A	C	A	−3.10	T	G	G	−8.77	T	−2.29
	A	T	G	−1.75	T	A	A	−11.16	.	.
	A	T	A	−12.42	C	A	G	−1.09	.	.
	C	C	G	−0.97	C	G	G	−13.28	.	.
	C	C	A	−20.55	C	A	A	−21.42	.	.
	C	T	G	−0.86	.	.	.	.	.	.
	C	T	A	−11.25	.	.	.	.	.	.
P1013										
	A	-	G	−2.00	T	-	G	−1.91	C	−11.55
	A	-	A	−5.44	T	-	A	−2.42	T	−2.30
	C	-	G	−0.36	C	-	G	−20.10	.	.
	C	-	A	−15.97	C	-	A	−21.44	.	.

Table 6.5. Weighted within-host selection coefficients for patients P5001-P5007. Selection coefficients (σ) are with respect to (potentially gapped) single-segment haplotypes defined by the variant loci in HA, NP, and PA. For example, any viral haplotype in patient P5001 exhibiting an A at position 545 and a G at position 1287 is under selective pressure by $\sigma = -2.09$.

Patient ID	HA				NP					PA		PB2	
	545	632	1287	w	264	372	913	1080	w	1704	w	912	w
P5001													
	A	-	G	−2.09	-	T	G	-	−5.60	C	−16.88	.	.
	A	-	A	−3.11	-	T	A	-	−12.15	T	−2.33	.	.
	C	-	G	−0.60	-	C	G	-	−21.35	.	.	.	.
	C	-	A	−21.17	-	C	A	-	−21.42	.	.	.	.
P5002													
	A	G	G	−1.40	-	T	G	-	−11.93	C	−17.74	.	.
	A	G	A	−20.14	-	T	A	-	−20.15	T	−1.06	.	.
	A	A	G	−1.76	-	C	G	-	−3.10	.	.	.	.
	A	A	A	−15.22	-	C	A	-	−19.18	.	.	.	.
	C	G	G	−0.86	.	.	.	.	.	.	.	.	.
	C	G	A	−16.59	.	.	.	.	.	.	.	.	.
	C	A	G	−0.06	.	.	.	.	.	.	.	.	.
	C	A	A	−16.85	.	.	.	.	.	.	.	.	.
P5004													
	A	-	G	−2.15	-	T	G	-	−9.09	C	−10.00	.	.
	A	-	A	−9.98	-	T	A	-	−5.79	T	−2.45	.	.
	C	-	G	−1.09	-	C	G	-	−1.16	.	.	.	.
	C	-	A	−20.26	-	C	A	-	−1.16	.	.	.	.
P5006													
	A	-	G	−1.68	-	T	G	-	−3.75	C	−6.35	.	.
	A	-	A	−5.83	-	T	A	-	−9.15	T	−2.42	.	.
	C	-	G	−1.06	-	C	G	-	−4.36	.	.	.	.
	C	-	A	−12.73	-	C	A	-	−21.92	.	.	.	.
P5007													
	A	-	G	−2.21	-	T	G	-	−8.34	C	−17.33	.	.
	A	-	A	−17.72	-	T	A	-	−20.32	T	−2.14	.	.
	C	-	G	−1.07	-	C	G	-	−13.35	.	.	.	.
	C	-	A	−21.01	-	C	A	-	−21.09	.	.	.	.

Table 6.6. Weighted within-host selection coefficients for patients P5018-P5021. Selection coefficients (σ) are with respect to (potentially gapped) single-segment haplotypes defined by the variant loci in HA, NP, and PA. For example, any viral haplotype in patient P5018 exhibiting an A at position 545 and a G at position 1287 is under selective pressure by $\sigma = -1.54$.

Patient ID	HA				NP					PA		PB2	
	545	632	1287	w	264	372	913	1080	w	1704	w	912	w
P5018													
	A	-	G	−1.54	A	T	G	C	−5.68	C	−9.51	.	.
	A	-	A	−5.93	A	T	A	C	−21.35	T	−2.21	.	.
	C	-	G	−1.00	A	T	G	T	−0.38	.	.	.	.
	C	-	A	−20.18	G	T	G	C	−0.56	.	.	.	.
	.	.	.	.	G	T	A	C	−6.59	.	.	.	.
	.	.	.	.	G	T	G	T	−0.38	.	.	.	.
	.	.	.	.	A	C	G	C	−4.02	.	.	.	.
	.	.	.	.	A	C	A	C	−22.16	.	.	.	.
	.	.	.	.	A	C	G	T	0.00	.	.	.	.
	.	.	.	.	G	C	G	C	−6.39	.	.	.	.
	.	.	.	.	G	C	A	C	0.00	.	.	.	.
	.	.	.	.	G	C	G	T	−6.25	.	.	.	.
P5019													
	A	-	G	−1.87	-	T	G	-	−10.16	C	−17.00	.	.
	A	-	A	−15.13	-	T	A	-	−6.60	T	−1.95	.	.
	C	-	G	−0.84	-	C	G	-	−12.24	.	.	.	.
	C	-	A	−20.66	-	C	A	-	−1.16	.	.	.	.
P5020													
	A	-	G	−1.94	-	T	G	-	−4.54	C	−12.18	G	−7.70
	A	-	A	−3.71	-	T	A	-	−3.73	T	−2.22	T	−7.56
	C	-	G	−1.09	-	C	G	-	−14.87	.	.	.	.
	C	-	A	−16.88	-	C	A	-	−21.42	.	.	.	.
P5021													
	A	-	G	−1.91	-	T	G	-	−4.79	C	−8.05	.	.
	A	-	A	−1.86	-	T	A	-	−20.07	T	−1.83	.	.
	C	-	G	−0.87	-	C	G	-	−2.00	.	.	.	.
	C	-	A	−9.18	-	C	A	-	−2.10	.	.	.	.

Table 6.7. Median inferred bottleneck sizes for the human challenge study accounting for within-host selection. Median values of $N^T > 100$ have been quoted as '100+', noting that the inference method is unable to properly distinguish bottlenecks of this magnitude.

Patient ID	Median bottleneck
1001	100+
1006	100+
1008	2
1010	2
1012	100+
1013	7
5001	2
5002	9
5004	11
5006	2
5007	2
5017	10
5018	33
5019	2
5020	2.5
5021	2

6.2.2.4 Transmission Inference

Transmission inference was performed for the 16 transmission events using the conservative C-value of $C = 138.86$. Inference methods accounting for and ignoring within-host selection were employed.

6.3 Results

6.3.1 Transmission Inference

Employing estimates of weighted selection to describe within-host selection where applicable (Tables 6.4 to 6.6), we inferred transmission bottlenecks both in the presence and absence of selection for within-host evolution. Using a model of selective neutrality with respect to transmission, outcomes from both methods are shown in Figures 6.1 to 6.3 in the form of overlapping histograms. Median bottleneck sizes are reported in Table 6.7.

Generally we find that accounting for within-host selection leads to transmission inferences with similar or larger bottlenecks than those obtained from

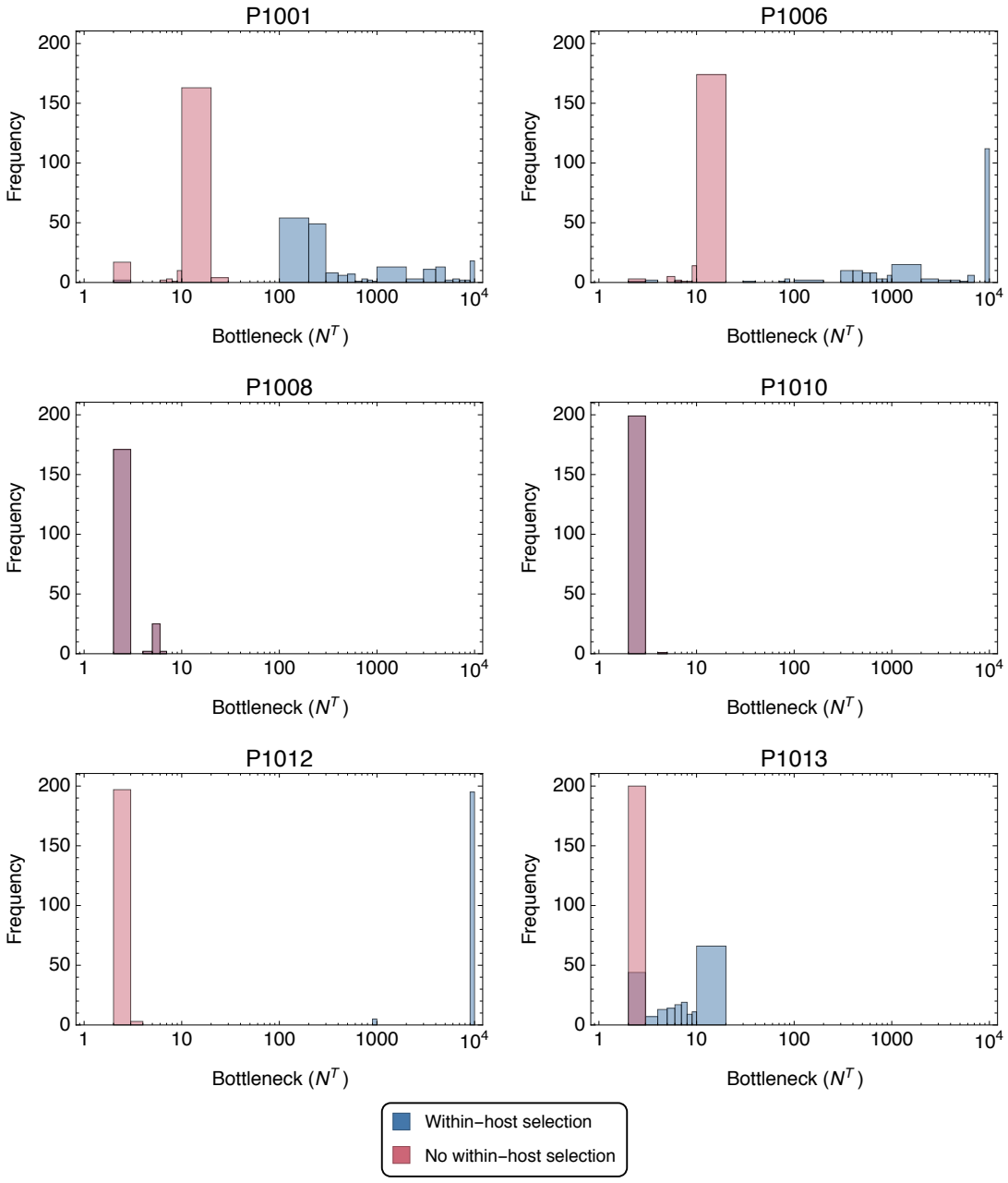


Figure 6.1. Histograms of bottleneck inferences for patients P1001–P1013 of the human challenge study. Inference methods either accounting for or ignoring within-host selection were employed with results displayed as overlapping histograms. Inferences were computed for a total of 200 analysis seeds. Outcomes were placed in bins with boundaries at $N^T = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10^4\}$.

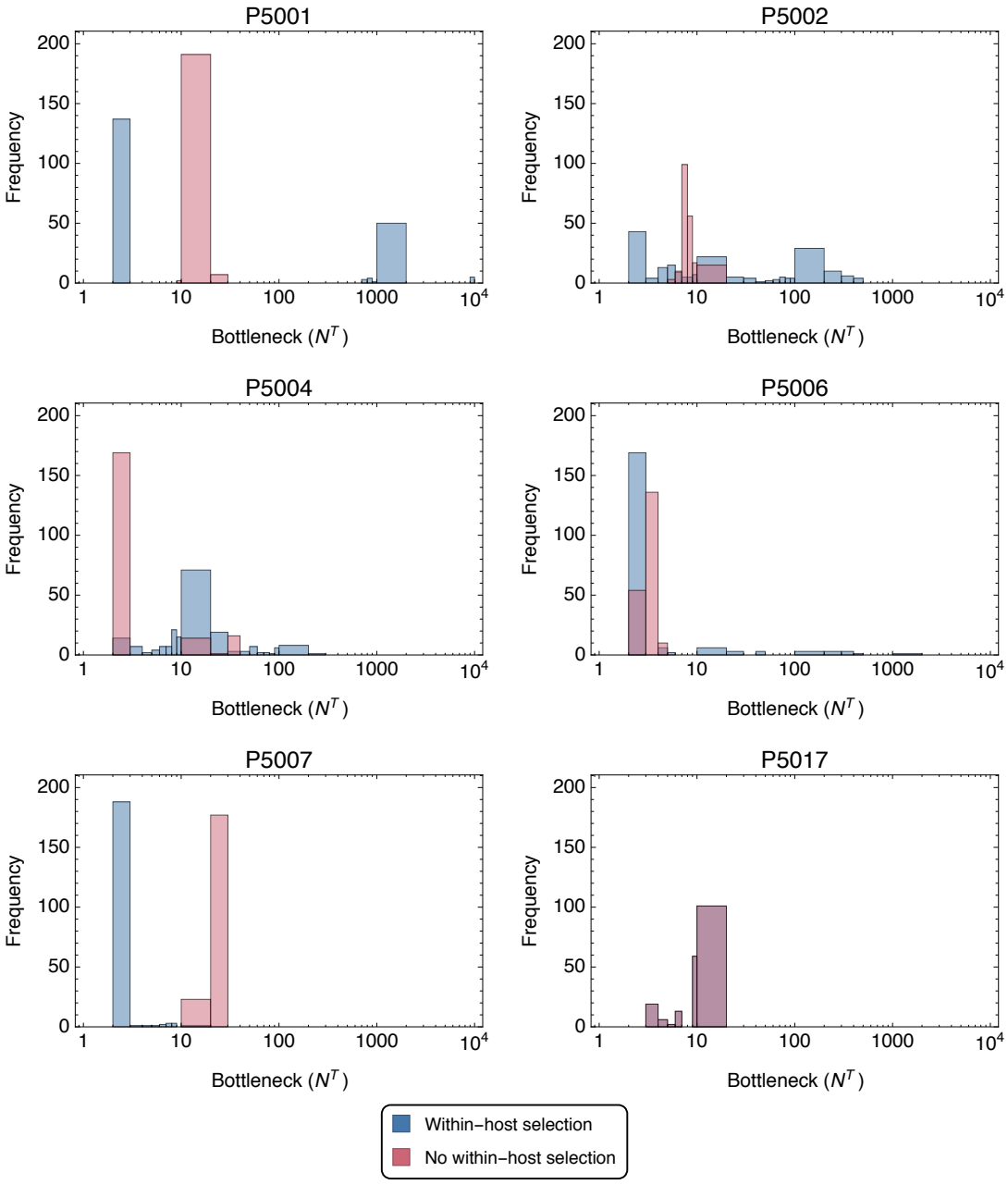


Figure 6.2. Histograms of bottleneck inferences for patients P5001–P5017 of the human challenge study. Inference methods either accounting for or ignoring within-host selection were employed with results displayed as overlapping histograms. Inferences were computed for a total of 200 analysis seeds. Outcomes were binned as described in the figure text of Figure 6.1.

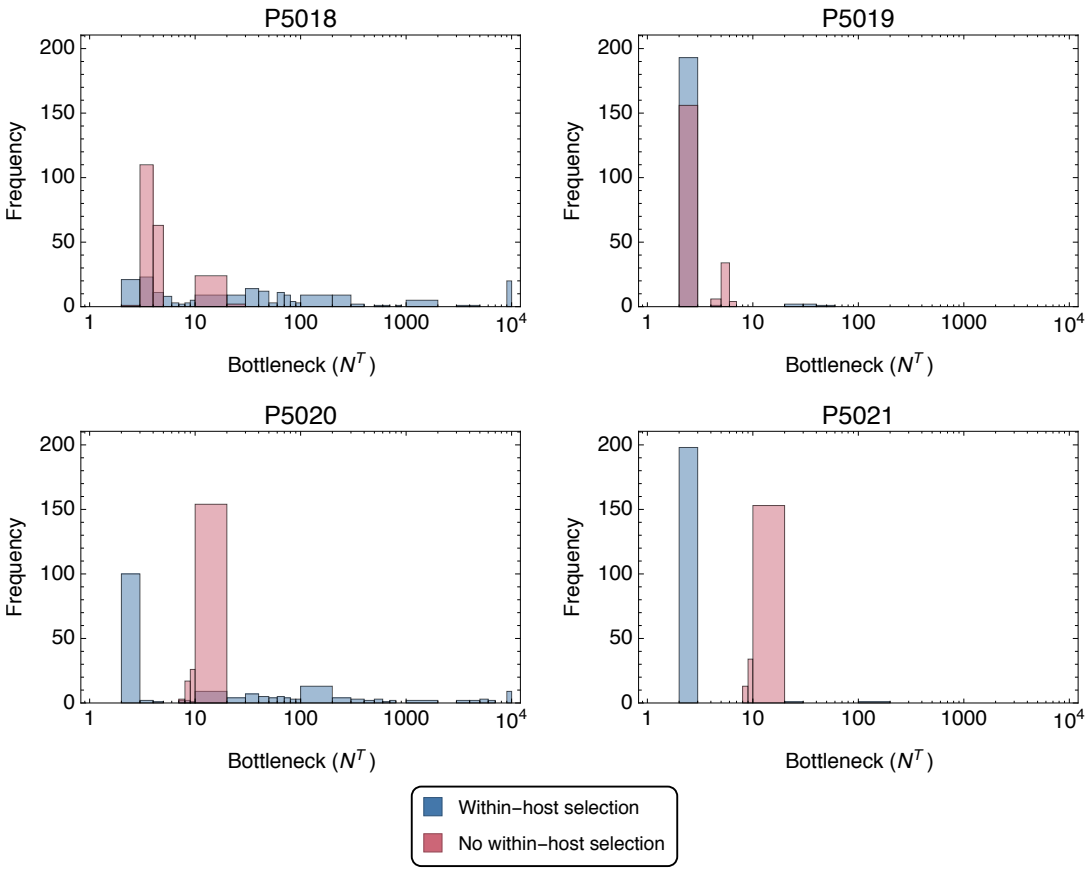


Figure 6.3. Histograms of bottleneck inferences for patients P5018–P5021 of the human challenge study. Inference methods either accounting for or ignoring within-host selection were employed with results displayed as overlapping histograms. Inferences were computed for a total of 200 analysis seeds. Outcomes were binned as described in the figure text of Figure 6.1.

inferences ignoring selection. Exceptions to this rule include individuals P5007 and P5021 for which the neutral model predicts the largest bottleneck sizes. Additionally, no within-host selection was inferred for patients P1008, P1010 and P5017, and, as a result, the two methods provided identical bottleneck inferences for these individuals. Collectively, bottleneck inferences in the absence of within-host selection exhibited median bottlenecks in the range of 2–22 virions whilst inferences in the presence of selection covered all possible values (as $N^T = 10^4$ was chosen as the maximum inferrable bottleneck size).

Inferences accounting for within-host selection showed a larger variability in estimated bottleneck sizes. This is especially true for bottleneck sizes of $N^T > 100$, inferred for the individuals P1001, P1006, and P1012, which may be explained by the model's inability to properly discriminate between bottleneck sizes of this magnitude. Using simulated data we have previously shown that estimation of bottleneck sizes in this regime leads to a considerable variance in inferences (Figure 3.11). This is due to increasing bottleneck sizes only bringing about marginal improvements in likelihoods, thus leaving the final bottleneck estimation highly dependent on the specific optimisation of $\boldsymbol{q}^B$ generated in the current statistical replicate. Whilst we are unlikely to feasibly differentiate between inferences of e.g. $N^T = 200$ and $N^T = 2000$, it is, nonetheless, fair to assume that these large estimated values arise from a true bottleneck of $N^T > 100$.

As experiment 1 involved inoculation with varying degrees of virus concentration, it is of relevance to consider whether there is a correlation between dosage size and founder population. Four inoculation dosages were used in the experiment: $10^{6.41}$ TCID_{50} (P1001), $10^{5.25}$ TCID_{50} (P1006, P1008), $10^{4.41}$ TCID_{50} (P1010, P1012), and $10^{3.08}$ TCID_{50} (P1013) (Moody et al. 2011). We here observe only a weak link between dosage and bottleneck size with both P1001 (high dose) and P1006 (medium high dose) having bottleneck inferences of $N^T > 100$, and P1010 (medium low dose) and P1013 (low dose) showing median bottlenecks of $N^T = 2$ and $N^T = 7$ respectively. However, neither P1008 (medium high dose, median $N^T = 2$) nor P1012 (medium low dose, median $N^T = 10^4$) fit well into this picture. Furthermore, all of the individuals in experiment 2 received an identical, but high, challenge dosage of 10^6 TCID_{50}, yet median bottleneck inferences vary between 2–33 virions.

Finally we observe a handful of transmissions resulting in bottleneck inferences of $N^T = 10^4$, i.e. the maximum inferrable bottleneck size under our model. Whilst in the case of the Brookes et al. data we generally inferred narrow bot-

tlenecks, we here observe much larger bottlenecks, making these inferences of $N^T = 10^4$ more plausible, e.g. in the case of P1001 and P1006. Regardless, in the case of P1012, accounting for within-host selection results in an unrealistically large bottleneck size; if we perform the analysis using a method ignoring the variance in $\boldsymbol{q}^B$ (see Section 5.5) we instead obtain an inference of $N^T = 1$. It has previously been noted that the data for individual P1012 have particularly low coverage in all but the MP and NS genes (Sobel Leonard et al. 2016), which may also contribute to an unrealistic bottleneck inference.

6.4 Discussion

In this chapter I introduced the concept of weighted selection and employed it for transmission inference in a human challenge study. Weighted selection represents an advancement of the principle of effective selection in cases where a multi-segment, multi-locus characterisation of selection exists. Accounting for within-host selection I here inferred a range of bottleneck values with multiple challenge subjects exhibiting founder populations of more than 100 virions. Conversely, not accounting for within-host selection lead to substantially lower estimates of bottleneck size with median bottlenecks here ranging from 2–33 virions.

The bottleneck sizes observed here are larger than what has previously been found in natural influenza infection in humans (McCrone et al. 2018). This suggests that infection by direct inoculation may differ in a quantitative manner from natural infections. Previous work in ferrets have utilised barcoded virus to demonstrate a significantly looser bottleneck associated with direct inoculation than with natural infection (Varble et al. 2014); our results here suggest that a similar effect may be found in humans. I note that where this previous calculation relied upon the use of a barcoded virus, our study has exploited the genetic variation naturally present in the virus; use of artificially engineered viruses, while providing a higher resolution picture of the transmission bottleneck, may be ethically challenging in human hosts.

The existence of higher bottlenecks in challenge studies than in natural infection would have consequences for what might be observed in the course of an experiment. A higher bottleneck increases the chances of minor variant alleles being preserved within the infecting population. Such alleles provide standing variation within the viral population, in turn supplying the material upon which selection can act. As such, phenomena such as resistance to antiviral drugs which

might not be observed in natural infections might more readily be observed in challenge studies if the corresponding alleles were to exist at low frequency in the viral inoculum. Such a phenomenon implies that challenge studies grant a conservative picture of what might happen during infection, undesirable events such as resistance having an increased likelihood of occurrence.

I note that the design of the specific study analysed here gave rise to an increased potential for evolutionary inference. In this study a human-adapted influenza virus was passaged through egg and cell culture before being administered to human hosts. During the period of growth in culture medium the virus gained in sequence diversity through the growth of adaptive mutations conferring culture-specific adaptation. Following the inoculation of human hosts, these variants were observed to revert, selection favouring the original human-adapted strain. This process of adaptation, occurring at a few loci within the genome, provided the genetic signals responsible for earlier inferences of reassortment rate and, in this case, inferences of the transmission bottleneck. Such a strategy, of inducing variation into the virus, might be of assistance in future studies of evolutionary dynamics within influenza infections.

While it is tempting to propose a potential correlation between inoculation dose and bottleneck size, our results were unable to confirm this. To further validate this would require a multitude of study subjects receiving a range of different inoculum dosages. Ultimately, whilst our method represents the best possible population genetic model for the data at hand, the only way to conclusively verify such a correlation would require the use of barcoded virus in human challenge studies, something which has yet to be seen.

I here obtain bottleneck inferences forming narrow and wide distributions and with estimates sampling the full range of inferrable bottleneck sizes. Such results, which are more diverse in nature than those obtained for the ferret experiment of a previous chapter, raise questions about the extent of the reliability of my inference method in this context. I note that the wide distributions observed for many of the individuals may be explained by an increasing degree of unspecificity as the bottleneck size increases. Furthermore, some of the extreme results may be explained by low coverage or lack of diversity, making inferences prone to error. Regardless, for a few of the inferences, e.g. P5001 and P5002, the method predicts either narrow or large bottlenecks, dependent upon the specific haplotype reconstruction arrived at by the method. Further work is needed to understand the underlying causes of these outcomes. As the method has been thoroughly tested using simulated data, any unusual outcomes are likely to arise

from a difference between the simulated data and the data obtained from an actual experiment. For example, sequencing errors may produce diversity artefacts which are difficult to capture using an overdispersed multinomial. Additionally, experimental data may derive from genotypes with very low frequencies, something which wasn't explored in simulations; this may impact the optimisation of $\boldsymbol{q}^B$. I note that gaining an increased understanding of the underlying optimisation procedure is challenging; the large dimensionality of the problem and the use of multiple evaluation spaces (both full and partial haplotype space) make it difficult to attribute a specific bottleneck inference to a certain genomic characteristic in the population. The possibility of an oversight in the code might also be considered, despite it proving robust to a broad range of previous testing. Additional software may be required to disentangle the different components of the model and to illuminate potential issues or options for improvement.

Finally I note that Sobel Leonard et al. (2016) predicted the presence of selective bottlenecks during their analysis of data from experiment 1 of the challenge study. In the previous chapter I was unable to reliably infer selection for increased transmissibility on the basis of potentially limited reassortment in the host animals. In this case, considering the direct inoculation of challenge subjects with stock virus, I have assumed a high rate of reassortment in the pre-transmission 'host', in line with experimental results from *in vitro* systems, combined with a low rate of reassortment in the human host post-transmission. Future work should aim to examine the possibility of selection for increased transmissibility in this dataset, while perhaps conducting further validation of the method's performance under an effective selection regime.

Chapter 7

Conclusion

In this thesis I have presented novel population genetic approaches to the inference of viral transmission events on the basis of short-read sequence data. Next-generation sequence data represents an inherent challenge to the interpretation of population dynamics, providing only limited insights into the underlying viral population. I first discussed approaches for the inference of full length viral haplotypes from sets of processed short-read data, something I, and previous authors, have referred to as partial haplotypes. I first considered an exhaustive approach to haplotype inference, using here a set of simple rules to construct a, generally large, set of haplotypes guaranteed to represent all the available partial haplotypes. I then demonstrated a degree of degeneracy in specifying a HIV population through maximum likelihood optimisation of haplotype frequencies; where an exhaustive set of haplotypes are used, multiple sets of inferred frequencies may be produced, each describing the partial haplotype observations equally well. Regardless, general trends were observed across optimisations and the method outperformed a random approach to frequency inference. Next, I employed haplotype reconstruction methods for the analysis of an influenza B time-series dataset, illustrating here the ability of haplotype methods to provide further insights than those available from conventional consensus sequence methods. The advantages of haplotype methods are likely to be of additional benefit in acute infections for which consensus sequence changes are minimal or non-existent. Finally, I presented a minimal approach to haplotype reconstruction, rooted in a simultaneous optimisation of haplotypes and frequencies, and demonstrated its ability to reduce the number of haplotypes given sparse partial haplotype data. Developing a maximum likelihood model of viral transmission, I illustrated biases inherent to single-locus variant ap-

proaches to transmission inference, here arising from a lack of accounting for within-segment linkage effects. Lastly, I employed the transmission model for the inference of bottleneck sizes from human influenza data, here obtaining generally narrow bottlenecks interspersed with occasionally larger inferences. This supports previous findings of narrow bottlenecks in humans, albeit suggesting infrequent exceptions to the rule.

Next, I present the statistical framework underlying my basic transmission model, which was used throughout the thesis. This framework takes a probabilistic approach, averaging over unobserved random variables resulting in a likelihood function for the variables of interest, here bottleneck size and extent of natural selection. The method is highly multi-dimensional, both at the level of full and partial haplotypes, its foundation being based on an exhaustive set of haplotypes. Evaluation using multivariate normal distributions ensures efficient computation, comparing with discrete methods, whilst accounting for between-haplotype correlations. I here thoroughly demonstrated the performance of the method for transmission inference, showing the ability of the method to jointly infer bottleneck size and extent of selection in transmission. I highlighted minor biases, such as the inherent underestimation of bottleneck sizes in the presence of noise, and, more importantly, identified severe biases arising from the neglect of accounting for noise or selection. I also discussed the challenges of model selection in biological data, here presenting a framework of adaptive BIC in which I employed a machine learning type approach for the evaluation of model selection penalties. Finally, I compared my model to the current state-of-the-art single-locus bottleneck inference algorithm of Sobel Leonard et al. Employing different growth factors, representing the rate at which the founder population expands, I obtained accurate bottleneck inferences for both the single- and multi-locus methods under neutrality and a growth factor of unity. Conversely, for increasing values of the growth factor, I observed substantial bottleneck overestimation under the single-locus method, something not found with my method. This demonstrates a high degree of flexibility within my framework for modelling a range of growth processes.

I then extended my model, allowing for a more realistic description of within-host growth processes, including the incorporation of selection for increased within-host adaptation. I here demonstrated that neglecting within-host selection resulted in bottleneck inference biases similar to those observed when neglecting selection for transmission; in order to separate the two effects of selection, I performed an independent estimation of the extent of within-host selection on

the basis of auxiliary time-series data. Considering drift, I observed that properly accounting for multiple rounds of within-host growth is important in the case of small growth factors, but has diminishing returns for a growth factor of 22, which is the default value employed in this thesis. Finally, I employed my transmission inference framework to an experimental study in ferrets, here accounting for within-host selection, resulting in narrow bottleneck inferences and no evidence of selection for increased transmissibility. Representing the first quantitative approach to bottleneck inference for this dataset, my results provide a qualitatively different interpretation of the data, noting that previous investigations predicted both narrow and loose bottlenecks, as well as the presence of selection.

Having established my transmission framework, I then explored limitations to my model and potential strategies for circumventing these. For the inference of selection I have previously assumed a large reassortment rate, allowing for the treatment of individual gene segments as independent. This is an important assumption, permitting the use of a single-segment haplotype model as opposed to a multi-segment approach; an across-segment model results in large numbers of potential haplotypes in all but the simplest cases, in turn making transmission inference infeasible. Large reassortment rates have been demonstrated in model animals, but have recently been found absent in humans. Erring on the side of caution, I here assumed a lack of reassortment in large mammals in general, including swine. Whilst the neutral model of my method could be applied in cases of low reassortment, I have previously identified strong biases in bottleneck inference from neglecting selection. As such, I developed an effective approach for the estimation of within-host selection, in which across-segment linkage effects renders selection time-dependent, and, for the task of within-host selection estimation, the effective selection is evaluated at the first two time points in the recipient. Producing an estimate for effective within-host selection, I applied my transmission model for the inference of bottleneck sizes in an experimental study on herds of swine. I here observed generally narrow bottlenecks in the range of 1–8 virions, but with occasional inferences as large as 57 transmitted viruses. This represents the first time a quantitative evaluation of transmission bottlenecks in pigs have been carried out. As the dataset is unspecific with respect to route of transmission, I computed bottleneck values for all transmission events that may potentially have taken place. Additionally, I also investigated methods for the inference of route of transmission networks, aiming to identify genuine transmission links in the dataset. I posited that large bottlenecks might be good

indicators of transmission, and thus identified the most likely transmission links in the dataset, of which only one transmission suggested a strong link. Comparing with existing methods of route of transmission inference, I found that estimations due to consensus and shared variant methods were either unspecific or prone to error due to narrow bottlenecks. Finally, I examined a rapid, sub-consensus minimum genetic distance approach, which showed good agreement with the results obtained from the full examination of transmission, suggesting that this might make for a useful preliminary investigation tool for route of transmission inference.

As a final frontier to transmission inference, I applied my model to an influenza transmission study in human challenge subjects. With model animals being limited in their ability to represent natural infection in humans, human challenge studies, wherein study subjects are directly inoculated with a pathogen, are of increasing value and use. Based on a multi-segment characterisation of a within-host fitness landscape, I presented an alternative approach for computing effective selection, here rooted in performing a weighted average across haplotypes. Assuming the multi-segment characterisation is well specified, this provides an accurate and mathematically aesthetic alternative to the effective selection approach mentioned above, not at least because the weighted approach is evaluated at a single time point only. Employing an estimate of within-host selection, I inferred a range of bottleneck sizes in the human challenge study, with multiple subjects exhibiting bottlenecks of more than a hundred virions. Whilst the majority of the bottlenecks are still narrow, I here observed a higher proportion of large bottlenecks than found by McCrone et al., suggesting that infection by direct inoculation differs quantitatively from natural infection. Despite some signatures, I was unable to confirm a correlation between inoculation dosage and bottleneck size; differently designed studies or the use of barcoded virus might be required to verify this.

Reflecting on the project as a whole, I note that my inference framework is significantly more advanced than existing methods, considering a range of confounding factors such as noise, multiple sources of selection and within-host growth; an important finding from this thesis is the identification of a range of sources that may potentially bias bottleneck inference. As shown here, it may not always be possible to account for all of these factors, for instance in cases where no auxiliary time-series data exits for the estimation of within-host selection. Rather than discounting methods not considering these confounding factors, I instead urge the scientific community to reflect on the possible biases

encountered when using simpler algorithms for bottleneck inference. Regardless, I note that even in cases where the more advanced aspects of my model may be inapplicable, the neutral version is still remarkably fast, not requiring adaptive BIC, providing a useful alternative to existing single-locus methods which do not account for haplotype structure and noise. Similarly, upon further development and scrutiny, our minimal haplotype transmission method is a likely candidate for becoming the state-of-the-art approach to neutral transmission.

My transmission model has a few drawbacks, primarily related to the exhaustive nature of the haplotype reconstruction method used within. When short-read data is sparse, the exhaustive method generates haplotypes in the thousands. As the method relies on repeated computation of matrix multiplications of dimension equal to the number of haplotypes, this is problematic. I here employed a filtering algorithm, removing haplotypes of increasing frequency until less than a hundred haplotypes remain. This number was chosen arbitrarily, although rooted in a substantial drop in computational efficiency when the number of haplotypes were in excess of this. Additionally, when haplotype reconstruction resulted in a large number of haplotypes, the optimisation of the pre-transmission population was subject to being underspecified; in the extreme case a total of 200 parameters needed fitting. Some of the spurious results reported here, e.g. unrealistically large bottleneck inferences, are likely to be due to a misrepresentation of the pre-transmission population. Finally, a large dimensionality might lead to errors of machine precision when a large array of small numbers are multiplied together. Ultimately, there is a need for a minimal approach to haplotype inference guaranteeing that all partial haplotypes are represented by a full haplotype; a requirement which our current minimal haplotype transmission method does not meet. Mathematically, this is known as the set cover problem, which is an NP-complete problem. There exists greedy algorithms for this problem, however, these aren't guaranteed to produce a set of haplotypes significantly smaller than those produced by our exhaustive haplotype reconstruction method. Conclusively, this may be a very difficult issue to address. Another important drawback is the adaptive BIC inference method. Whilst integral for accurate inference of selection, the optimisation of BIC penalties is a time consuming process as it requires the generation of a significant amount of simulated data.

Overall, I have identified generally narrow bottlenecks in influenza transmission events, suggesting that this presents a general picture in mammals. From simulated data I noted a difficulty in identifying selection for transmission in

the presence of narrow bottlenecks. Given estimates of bottlenecks generally less than ten virions, it is likely that inference of selection in natural influenza transmission is close to being an impossible task. I note that there is a potential for inferring selection in the human challenge study, for which a handful of bottlenecks were of considerable magnitude. Future work should explore this possibility.

I have developed two measures of effective selection, aiming to estimate within-host selection in the presence of reassortment. I note that, whilst mathematically sound, there may be cases where these inferences cannot be trusted. In particular, the methods rely on approximate constancy of linkage disequilibrium during short time intervals. Given that I evaluate effective selection at the first available time points after transmission, it is possible that this doesn't correspond well with the selection present at the founding of the infection. Regardless, I believe the currently proposed methods provide the best possible approach to an otherwise impossible problem. In conclusion, bottleneck inferences using effective selection should be considered as a whole, rather than by taking individual inferences at face value.

Finally I note that my method lacks the ability to account for mutation. Currently I filter away partial haplotypes observed only after transmission, considering these to have arisen *de novo*. Whilst generally unlikely to impact the bottleneck inference appreciably, future work could consider means of incorporating mutation into the framework. Lastly, I note that whilst my method has been developed for the inference of influenza transmission events, the majority of the algorithm is general in nature. The main aspect linking the project to flu transmission is the assumption of a segmented genome, from which inference power arises. Potentially, given appreciable homologous recombination capable of breaking up linkage, this assumption could be reformulated for different viral species. At the very core of my method lies the compound distribution approach. I consider this to be an extremely powerful method and I believe that it warrants further use in studies where marginalisation over unknown quantities is a key objective.

Bibliography

Abel, S. et al. (2015). 'Analysis of Bottlenecks in Experimental Models of Infection'. In: *PLoS Pathogens* vol. 11, no. 6, e1004823–7.

Acevedo, A., L. Brodsky and R. Andino (2014). 'Mutational and fitness landscapes of an RNA virus revealed through population sequencing.' In: *Nature* vol. 505, no. 7485, pp. 686–690.

Achaz, G. et al. (2014). 'The reproducibility of adaptation in the light of experimental evolution with whole genome sequencing.' In: *Adv Exp Med Biol.* Vol. 781, pp. 211–31.

Aggarwal, C. C., A. Hinneburg and D. A. Keim (2001). 'On the Surprising Behavior of Distance Metrics in High Dimensional Space'. In: Springer Berlin Heidelberg, pp. 420–434.

Akaike, H. (1974). 'A new look at the statistical model identification'. In: *IEEE Transactions on Automatic Control* vol. 19, no. 6, pp. 716–723.

Baccam, P. et al. (2006). 'Kinetics of influenza A virus infection in humans.' In: *Journal of virology* vol. 80, no. 15, pp. 7590–9.

Bambery, B. et al. (2016). 'Ethical Criteria for Human Challenge Studies in Infectious Diseases'. In: *Public Health Ethics* vol. 9, no. 1, pp. 92–103.

Barton, J. P. et al. (2016). 'Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable'. In: *Nature Communications* vol. 7, no. 1, p. 11660.

Beerenwinkel, N. and O. Zagordi (2011). 'Ultra-deep sequencing for the analysis of viral populations'. In: *Current Opinion in Virology* vol. 1, no. 5, pp. 413–418.

Belser, J. A., J. M. Katz and T. M. Tumpey (2011). 'The ferret as a model organism to study influenza A virus infection.' In: *Disease Models & Mechanisms* vol. 4, no. 5, pp. 575–9.

Bergstrom, C. T., P. McElhany and L. A. Real (1999). 'Transmission bottlenecks as determinants of virulence in rapidly evolving pathogens.' In: *Proceedings of the National Academy of Sciences of the United States of America* vol. 96, no. 9, pp. 5095–5100.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag. ISBN: 0387310738.

Blanquart, F. et al. (2016). 'A transmission-virulence evolutionary trade-off explains attenuation of HIV-1 in Uganda.' In: *eLife* vol. 5, p. 2171.

Bollback, J. P., T. L. York and R. Nielsen (2008). 'Estimation of 2Nes From Temporal Allele Frequency Data'. In: *Genetics* vol. 179, no. 1, pp. 497–502.

Boni, M. F. et al. (2008). 'Homologous recombination is very rare or absent in human influenza A virus.' In: *Journal of Virology* vol. 82, no. 10, pp. 4807–11.

Bouckaert, R. et al. (2014). 'BEAST 2: A Software Platform for Bayesian Evolutionary Analysis'. In: *PLOS Computational Biology* vol. 10, no. 4, pp. 1–6.

Bowman, A. S. et al. (2017). 'Influenza A(H3N2) Virus in Swine at Agricultural Fairs and Transmission to Humans, Michigan and Ohio, USA, 2016'. In: *Emerging Infectious Diseases* vol. 23, no. 9, pp. 1551–1555.

Breban, R., J. Riou and A. Fontanet (2013). 'Interhuman transmissibility of Middle East respiratory syndrome coronavirus: estimation of pandemic risk'. In: *The Lancet* vol. 382, no. 9893, pp. 694–699.

Brookes, S. et al. (2010). *Unpublished Porcine Transmission Study.*

Buhnerkempe, M. G. et al. (2015). 'Mapping influenza transmission in the ferret model to transmission in humans'. In: *eLife* vol. 4.

Bull, J. J., M. R. Badgett and H. A. Wichman (2000). 'Big-Benefit Mutations in a Bacteriophage Inhibited with Heat'. In: *Molecular Biology and Evolution* vol. 17, no. 6, pp. 942–950.

Carrat, F. and A. Flahault (2007). 'Influenza vaccine: The challenge of antigenic drift'. In: *Vaccine* vol. 25, no. 39-40, pp. 6852–6862.

Carrat, F. et al. (2008). 'Time Lines of Infection and Disease in Human Influenza: A Review of Volunteer Challenge Studies'. In: *American Journal of Epidemiology* vol. 167, no. 7, pp. 775–785.

Cauldwell, A. V. et al. (2014). 'Viral determinants of influenza A virus host range'. In: *Journal of General Virology* vol. 95, no. 6, pp. 1193–1210.

Centers for Disease Control and Prevention (2018). *Types of Influenza Viruses*. Accessed: 08-01-2019. URL: `https://www.cdc.gov/flu/about/viruses/types.htm`.

Chare, E. R., E. A. Gould and E. C. Holmes (2003). 'Phylogenetic analysis reveals a low rate of homologous recombination in negative-sense RNA viruses'. In: *Journal of General Virology* vol. 84, no. 10, pp. 2691–2703.

Charlesworth, B. (2009). 'Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation'. In: *Nature Reviews Genetics* vol. 10, no. 3, pp. 195–205.

Charu, V. et al. (2017). 'Human mobility and the spatial transmission of influenza in the United States'. In: *PLOS Computational Biology* vol. 13, no. 2, e1005382.

Chen, X. et al. (2018). 'Host Immune Response to Influenza A Virus Infection.' In: *Frontiers in immunology* vol. 9, p. 320.

Coombs, D., M. A. Gilchrist and C. L. Ball (2007). 'Evaluating the importance of within- and between-host selection pressures on the evolution of chronic pathogens'. In: *Theoretical Population Biology* vol. 72, no. 4, pp. 576–591.

Cowling, B. J. et al. (2013). 'Aerosol transmission is an important mode of influenza A virus spread'. In: *Nature Communications* vol. 4, p. 1657.

Crawford, P. C. et al. (2005). 'Transmission of Equine Influenza Virus to Dogs'. In: *Science* vol. 310, no. 5747, pp. 482–485.

Darton, T. C. et al. (2015). 'Design, recruitment, and microbiological considerations in human challenge studies'. In: *The Lancet Infectious Diseases* vol. 15, no. 7, pp. 840–851.

Derdeyn, C. A. et al. (2004). 'Envelope-Constrained Neutralization-Sensitive HIV-1 after Heterosexual Transmission'. In: *Science* vol. 303, no. 5666, pp. 2019–2022.

Dijk, E. L. van et al. (2014). 'Ten years of next-generation sequencing technology.' In: *Trends in Genetics* vol. 30, no. 9, pp. 418–426.

Dinis, J. M. et al. (2016). 'Deep Sequencing Reveals Potential Antigenic Variants at Low Frequencies in Influenza A Virus-Infected Humans.' In: *Journal of Virology* vol. 90, no. 7, pp. 3355–3365.

Doud, M. B., J. M. Lee and J. D. Bloom (2018). 'How single mutations affect viral escape from broad and narrow antibodies to H1 influenza hemagglutinin'. In: *Nature Communications* vol. 9, no. 1, p. 1386.

Edwards, C. T. T. et al. (2006). 'Population genetic estimation of the loss of genetic diversity during horizontal transmission of HIV-1.' In: *BMC Evolutionary Biology* vol. 6, p. 28.

Feder, A. F., S. Kryazhimskiy and J. B. Plotkin (2014). 'Identifying signatures of selection in genetic time series.' In: *Genetics* vol. 196, no. 2, pp. 509–522.

Felsenstein, J. (1971). 'Inbreeding and variance effective numbers in populations with overlapping generations.' In: *Genetics* vol. 68, no. 4, pp. 581–597.

Ferguson, A. L. et al. (2013). 'Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design.' In: *Immunity* vol. 38, no. 3, pp. 606–17.

Fischer, A. et al. (2014). 'High-Definition Reconstruction of Clonal Composition in Cancer'. In: *Cell Reports* vol. 7, no. 5, pp. 1740–1752.

Foll, M. et al. (2014). 'Influenza Virus Drug Resistance: A Time-Sampled Population Genetics Perspective'. In: *PLoS Genetics* vol. 10, no. 2, e1004185.

Fraser, C. et al. (2009). 'Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings'. In: *Science* vol. 324, no. 5934, pp. 1557–1561.

Frise, R. et al. (2016). 'Contact transmission of influenza virus between ferrets imposes a looser bottleneck than respiratory droplet transmission allowing propagation of antiviral resistance.' In: *Scientific Reports* vol. 6, no. 1, p. 29793.

Garten, R. J. et al. (2009). 'Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans.' In: *Science* vol. 325, no. 5937, pp. 197–201.

Geoghegan, J. L. et al. (2018). 'Continental synchronicity of human influenza virus epidemics despite climactic variation'. In: *PLOS Pathogens* vol. 14, no. 1, e1006780.

Geyer, C. J. (2006). *Stat 5101 Notes: Brand Name Distributions*. Accessed: 19-10-2018. URL: `http://www.stat.umn.edu/geyer/old06/5101/notes/brand.pdf`.

Griffiths, A. J. F. et al. (2000). *An Introduction to Genetic Analysis. 7th edition.* Freeman, New York.

Gustin, K. M. et al. (2011). 'Influenza virus aerosol exposure and analytical system for ferrets.' In: *Proceedings of the National Academy of Sciences of the United States of America* vol. 108, no. 20, pp. 8432–7.

Gutiérrez, S., Y. Michalakis and S. Blanc (2012). 'Virus population bottlenecks during within-host progression and host-to-host transmission'. In: *Current Opinion in Virology* vol. 2, no. 5, pp. 546–555.

Hamada, N. et al. (2012). 'Intrahost emergent dynamics of oseltamivir-resistant virus of pandemic influenza A (H1N1) 2009 in a fatally immunocompromised patient'. In: *Journal of Infection and Chemotherapy* vol. 18, no. 6, pp. 865–871.

Hasegawa, M., H. Kishino and T. Yano (1985). 'Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.' In: *Journal of Molecular Evolution* vol. 22, no. 2, pp. 160–74.

Haß, J. et al. (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'. In: *Infection, Genetics and Evolution* vol. 11, no. 2, pp. 437–441.

Herfst, S. et al. (2012). 'Airborne transmission of influenza A/H5N1 virus between ferrets.' In: *Science* vol. 336, no. 6088, pp. 1534–41.

Hirve, S. et al. (2016). 'Influenza Seasonality in the Tropics and Subtropics – When to Vaccinate?' In: *PLOS ONE* vol. 11, no. 4, e0153003.

Huang, Y. et al. (2011). 'Temporal Dynamics of Host Molecular Responses Differentiate Symptomatic and Asymptomatic Influenza A Infection'. In: *PLOS Genetics* vol. 7, no. 8, e1002234.

Humphrey, W., A. Dalke and K. Schulten (1996). 'VMD – Visual Molecular Dynamics'. In: *Journal of Molecular Graphics* vol. 14, pp. 33–38.

Hurt, A. C. et al. (2009). 'Oseltamivir resistance and the H274Y neuraminidase mutation in seasonal, pandemic and highly pathogenic influenza viruses.' In: *Drugs* vol. 69, no. 18, pp. 2523–2531.

Illingworth, C. J. R. (2015). 'Fitness Inference from Short-Read Data: Within-Host Evolution of a Reassortant H5N1 Influenza Virus.' In: *Molecular Biology and Evolution* vol. 32, no. 11, pp. 3012–26.

— (2016). 'SAMFIRE: Multi-locus variant calling for time-resolved sequence data'. In: *Bioinformatics* vol. 32, no. 14, pp. 2208–2209.

Illingworth, C. J. R., A. Fischer and V. Mustonen (2014). 'Identifying selection in the within-host evolution of influenza using viral sequence data.' In: *PLOS Computational Biology* vol. 10, no. 7.

Illingworth, C. J. R. and V. Mustonen (2011). 'Distinguishing driver and passenger mutations in an evolutionary history categorized by interference.' In: *Genetics* vol. 189, no. 3, pp. 989–1000.

— (2012a). 'A method to infer positive selection from marker dynamics in an asexual population.' In: *Bioinformatics* vol. 28, no. 6, pp. 831–7.

— (2012b). 'Components of Selection in the Evolution of the Influenza Virus: Linkage Effects Beat Inherent Selection'. In: *PLOS Pathogens* vol. 8, no. 12, e1003091.

Illingworth, C. J. R. et al. (2017). 'On the effective depth of viral sequence data.' In: *Virus Evolution* vol. 3, no. 2, vex030.

Imai, M. et al. (2012). 'Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets'. In: *Nature* vol. 486, no. 7403, pp. 420–428.

Ince, W. L. et al. (2013). 'Reassortment Complements Spontaneous Mutation in Influenza A Virus NP and M1 Genes To Accelerate Adaptation to a New Host'. In: *Journal of Virology* vol. 87, no. 8, pp. 4330–4338.

Iyer, S. et al. (2015). 'Comparison of Major and Minor Viral SNPs Identified through Single Template Sequencing and Pyrosequencing in Acute HIV-1 Infection.' In: *PLOS ONE* vol. 10, no. 8, e0135903.

Jin, C. et al. (2017). 'Efficacy and immunogenicity of a Vi-tetanus toxoid conjugate vaccine in the prevention of typhoid fever using a controlled human infection model of Salmonella Typhi: a randomised controlled, phase 2b trial.' In: *Lancet* vol. 390, no. 10111, pp. 2472–2480.

Johnson, N. P. A. S. et al. (2002). 'Updating the Accounts: Global Mortality of the 1918-1920 "Spanish" Influenza Pandemic'. In: *Bulletin of the History of Medicine* vol. 76, no. 1, pp. 105–115.

Johnson, S. G. (n.d.). *Multi-dimensional adaptive integration (cubature) in C*. Accessed: 19-01-2018. URL: `https://github.com/stevengj/cubature`.

Kass, R. E. and A. E. Raftery (1995). 'Bayes factors'. In: *Journal of the American Statistical Association* vol. 90, no. 430, pp. 773–795.

Kawaoka, Y. et al. (1998). 'Molecular Basis for the Generation in Pigs of Influenza A Viruses with Pandemic Potential'. In: *Journal of Virology* vol. 72, no. 9, pp. 7367–7373.

Khiabanian, H. et al. (2015). 'High-resolution Genomic Surveillance of 2014 Ebolavirus Using Shared Subclonal Variants'. In: *PLOS Currents* vol. 7, pp. 1–17.

Killingley, B. et al. (2012). 'Use of a Human Influenza Challenge Model to Assess Person-to-Person Transmission: Proof-of-Concept Study.' In: *The Journal of Infectious Diseases* vol. 205, no. 1, pp. 35–43.

Kimura, M. (1955). 'Stochastic processes and distribution of gene frequencies under natural selection.' In: *Cold Spring Harbor Symposia on Quantitative Biology* vol. 20, pp. 33–53.

Koch, R. (1876). 'The etiology of anthrax, based on the life history of Bacillus anthracis'. In: *Beitrage zur Biologie der Pflanzen* vol. 2, no. 2, 277–310 (in German).

— (1984). 'The etiology of tuberculosis'. In: *Mitteilungen aus dem Kaiserlichen Gesundheitsamte* vol. 2, 1–88 (in German).

Koel, B. F. et al. (2013). 'Substitutions Near the Receptor Binding Site Determine Major Antigenic Change During Influenza Virus Evolution'. In: *Science* vol. 342, no. 6161, pp. 976–979.

Koelle, K. and D. A. Rasmussen (2015). 'The effects of a deleterious mutation load on patterns of influenza A/H3N2's antigenic evolution in humans.' In: *eLife* vol. 4, e07361.

Kofler, R. and C. Schlötterer (2014). 'A guide for the design of evolve and resequencing studies.' In: *Molecular Biology and Evolution* vol. 31, no. 2, pp. 474–483.

Krimbas, C. B. and S. Tsakas (1971). 'The Genetics of Dacus Oleae. V. Changes of esterase polymorphism in a natural population following insecticide control—selection or drift?' In: *Evolution* vol. 25, no. 3, pp. 454–460.

Kuiken, T. et al. (2006). 'Host species barriers to influenza virus infections.' In: *Science* vol. 312, no. 5772, pp. 394–397.

Kumar, S., G. Stecher and K. Tamura (2016). 'MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets'. In: *Molecular Biology and Evolution* vol. 33, no. 7, pp. 1870–1874.

Lacerda, M. and C. Seoighe (2014). 'Population genetics inference for longitudinally-sampled mutants under strong selection.' In: *Genetics* vol. 198, no. 3, pp. 1237–50.

Laehnemann, D., A. Borkhardt and A. C. McHardy (2016). 'Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction.' In: *Briefings in Bioinformatics* vol. 17, no. 1, pp. 154–179.

Leitner, T. et al. (1996). 'Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis.' In: *Proceedings of the National Academy of Sciences of the United States of America* vol. 93, no. 20, pp. 10864–9.

Lemey, P. et al. (2014). 'Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2'. In: *PLOS Pathogens* vol. 10, no. 2, e1003932.

Li, H. et al. (2009). 'The Sequence Alignment/Map format and SAMtools.' In: *Bioinformatics* vol. 25, no. 16, pp. 2078–9.

Li, L. M., N. C. Grassly and C. Fraser (2017). 'Quantifying Transmission Heterogeneity Using Both Pathogen Phylogenies and Incidence Time Series'. In: *Molecular Biology and Evolution* vol. 34, no. 11, pp. 2982–2995.

Lieberman, T. D. et al. (2014). 'Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures'. In: *Nature Genetics* vol. 46, no. 1, pp. 82–87.

Linster, M. et al. (2014). 'Identification, Characterization, and Natural Selection of Mutations Driving Airborne Transmission of A/H5N1 Virus'. In: *Cell* vol. 157, no. 2, pp. 329–339.

Lipsitch, M. et al. (2016). 'Viral factors in influenza pandemic risk assessment.' In: *eLife* vol. 5, 316ra192.

Liu, S. et al. (2009). 'Panorama Phylogenetic Diversity and Distribution of Type A Influenza Virus'. In: *PLOS ONE* vol. 4, no. 3, e5022.

Louie, R. H. Y. et al. (2018). 'Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies.' In: *Proceedings of the National Academy of Sciences of the United States of America* vol. 115, no. 4, E564–E573.

Łuksza, M. and M. Lässig (2014). 'A predictive fitness model for influenza'. In: *Nature* vol. 507, no. 7490, pp. 57–61.

Lumby, C. K., N. R. Nene and C. J. R. Illingworth (2018). 'A novel framework for inferring parameters of transmission from viral sequence data'. In: *PLOS Genetics* vol. 14, no. 10, e1007718.

Lythgoe, K. A. et al. (2017). 'Short-Sighted Virus Evolution and a Germline Hypothesis for Chronic Viral Infections'. In: *Trends in Microbiology* vol. 25, no. 5, pp. 336–348.

Ma, W., R. E. Kahn and J. A. Richt (2008). 'The pig as a mixing vessel for influenza viruses: Human and veterinary implications.' In: *Journal of Molecular and Genetic Medicine* vol. 3, no. 1, pp. 158–66.

Malaspinas, A.-S. et al. (2012). 'Estimating allele age and selection coefficient from time-serial data.' In: *Genetics* vol. 192, no. 2, pp. 599–607.

Marshall, N. et al. (2013). 'Influenza Virus Reassortment Occurs with High Frequency in the Absence of Segment Mismatch'. In: *PLOS Pathogens* vol. 9, no. 6, e1003421.

McArthur, M. and M. A. Shirley (2011). 'The utility of human challenge studies in vaccine development: lessons learned from cholera'. In: *Vaccine: Development and Therapy* vol. 1, p. 3.

McClain, M. T. et al. (2016). 'A genomic signature of influenza infection shows potential for presymptomatic detection, guiding early therapy, and monitoring clinical responses'. In: *Open Forum Infectious Diseases* vol. 3, no. 1, ofw007.

McCrone, J. T. and A. S. Lauring (2016). 'Measurements of Intrahost Viral Diversity Are Extremely Sensitive to Systematic Errors in Variant Calling.' In: *Journal of Virology* vol. 90, no. 15, pp. 6884–6895.

McCrone, J. T. et al. (2018). 'Stochastic processes constrain the within and between host evolution of influenza virus'. In: *eLife* vol. 7, e35962.

Miao, H. et al. (2010). 'Quantifying the Early Immune Response and Adaptive Immune Response Kinetics in Mice Infected with Influenza A Virus.' In: *Journal of Virology* vol. 84, no. 13, pp. 6687–98.

Moncla, L. H. et al. (2016). 'Selective Bottlenecks Shape Evolutionary Pathways Taken during Mammalian Adaptation of a 1918-like Avian Influenza Virus'. In: *Cell Host & Microbe* vol. 19, no. 2, pp. 169–180.

Monsion, B. et al. (2008). 'Large Bottleneck Size in Cauliflower Mosaic Virus Populations during Host Plant Colonization'. In: *PLOS Pathogens* vol. 4, no. 10, e1000174.

Moody, M. A. et al. (2011). 'H3N2 Influenza Infection Elicits More Cross-Reactive and Less Clonally Expanded Anti-Hemagglutinin Antibodies Than Influenza Vaccination'. In: *PLOS ONE* vol. 6, no. 10, e25797.

Mosimann, J. E. (1962). 'On the Compound Multinomial Distribution, the Multivariate β-Distribution, and Correlations Among Proportions'. In: *Biometrika* vol. 49, no. 1/2, p. 65.

Murcia, P. R. et al. (2010). 'Intra- and Interhost Evolutionary Dynamics of Equine Influenza Virus'. In: *Journal of Virology* vol. 84, no. 14, pp. 6943–6954.

Murcia, P. R. et al. (2012). 'Evolution of an Eurasian Avian-like Influenza Virus in Naïve and Vaccinated Pigs.' In: *PLOS pathogens* vol. 8, no. 5, e1002730.

Nature (2008). 'The long war against flu'. In: *Nature* vol. 454, no. 7201, pp. 137–137.

Neher, R. A. and B. I. Shraiman (2009). 'Competition between recombination and epistasis can cause a transition from allele to genotype selection.' In: *Proceedings of the National Academy of Sciences* vol. 106, no. 16, pp. 6866–6871.

Nishiura, H., H.-L. Yen and B. J. Cowling (2013). 'Sample Size Considerations for One-to-One Animal Transmission Studies of the Influenza A Viruses'. In: *PLoS ONE* vol. 8, no. 1, e55358–7.

Oh, D. Y. et al. (2018). 'Selection of multi-drug resistant influenza A and B viruses under zanamivir pressure and their replication fitness in ferrets.' In: *Antiviral Therapy* vol. 23, no. 4, pp. 295–306.

O'Hara, R. B. (2005). 'Comparing the effects of genetic drift and fluctuating selection on genotype frequency changes in the scarlet tiger moth'. In: *Proceedings of the Royal Society B: Biological Sciences* vol. 272, no. 1559, pp. 211–217.

Pagani, L. et al. (2015). 'Transmission and effect of multiple clusters of seasonal influenza in a Swiss geriatric hospital.' In: *Journal of the American Geriatrics Society* vol. 63, no. 4, pp. 739–44.

Palese, P. and T. T. Wang (2012). 'H5N1 influenza viruses: facts, not fear.' In: *Proceedings of the National Academy of Sciences* vol. 109, no. 7, pp. 2211–2213.

Pauly, M. D., M. C. Procario and A. S. Lauring (2017). 'A novel twelve class fluctuation test reveals higher than expected mutation rates for influenza A viruses.' In: *eLife* vol. 6, p. 686.

Pawelek, K. A. et al. (2012). 'Modeling Within-Host Dynamics of Influenza Virus Infection Including Immune Responses'. In: *PLOS Computational Biology* vol. 8, no. 6.

Peng, J. et al. (2014). 'The Origin of Novel Avian Influenza A (H7N9) and Mutation Dynamics for Its Human-To-Human Transmissible Capacity'. In: *PlOS one* vol. 9, no. 3, e93094.

Poon, L. L. M. et al. (2016). 'Quantifying influenza virus diversity and transmission in humans'. In: *Nature Genetics* vol. 48, no. 2, pp. 195–200.

Raghwani, J. et al. (2018). 'Evolution of HIV-1 within untreated individuals and at the population scale in Uganda'. In: *PLOS Pathogens* vol. 14, no. 7, e1007167.

Rambaut, A. (n.d.). *FigTree*. Accessed: 08-06-2019. URL: `http://tree.bio.ed.ac.uk/software/figtree/`.

Riedel, S. (2005). 'Edward Jenner and the history of smallpox and vaccination.' In: *Proceedings (Baylor University. Medical Center)* vol. 18, no. 1, pp. 21–5.

Rouzine, I. M., A. Rodrigo and J. M. Coffin (2001). 'Transition between stochastic evolution and deterministic evolution in the presence of selection: general theory and application to virology'. In: *Microbiology and Molecular Biology Reviews* vol. 65, no. 1, pp. 151–185.

Russell, C. A. et al. (2008). 'The global circulation of seasonal influenza A (H3N2) viruses.' In: *Science* vol. 320, no. 5874, pp. 340–6.

Russell, C. A. et al. (2012). 'The potential for respiratory droplet-transmissible A/H5N1 influenza virus to evolve in a mammalian host.' In: *Science* vol. 336, no. 6088, pp. 1541–7.

Sacristan, S. et al. (2003). 'Estimation of Population Bottlenecks during Systemic Movement of Tobacco Mosaic Virus in Tobacco Plants'. In: *Journal of Virology* vol. 77, no. 18, pp. 9906–9911.

Sandmann, S. et al. (2017). 'Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data.' In: *Scientific Reports* vol. 7, p. 43169.

Sandt, C. E. van de, J. H. C. M. Kreijtz and G. F. Rimmelzwaan (2012). 'Evasion of influenza A viruses from innate and adaptive immune responses.' In: *Viruses* vol. 4, no. 9, pp. 1438–76.

Sautto, G. A., G. A. Kirchenbaum and T. M. Ross (2018). 'Towards a universal influenza vaccine: different approaches for one goal'. In: *Virology Journal* vol. 15, no. 1, p. 17.

Scally, A. and R. Durbin (2012). 'Revising the human mutation rate: implications for understanding human evolution.' In: *Nature Reviews Genetics* vol. 13, no. 10, pp. 745–753.

Schwarz, G. (1978). 'Estimating the Dimension of a Model'. In: *The Annals of Statistics* vol. 6, no. 2, pp. 461–464.

Scoizec, A. et al. (2018). 'Airborne Detection of H5N8 Highly Pathogenic Avian Influenza Virus Genome in Poultry Farms, France'. In: *Frontiers in Veterinary Science* vol. 5, no. Feb.

Shaman, J. and M. Kohn (2009). 'Absolute humidity modulates influenza survival, transmission, and seasonality.' In: *Proceedings of the National Academy of Sciences of the United States of America* vol. 106, no. 9, pp. 3243–8.

Shinya, K. et al. (2006). 'Influenza virus receptors in the human airway'. In: *Nature* vol. 440, no. 7083, pp. 435–436.

Shortridge, K. F. et al. (1977). 'Persistence of Hong Kong influenza virus variants in pigs.' In: *Science* vol. 196, no. 4297, pp. 1454–5.

Sidorenko, Y. and U. Reichl (2004). 'Structured model of influenza virus replication in MDCK cells'. In: *Biotechnology and Bioengineering* vol. 88, no. 1, pp. 1–14.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall, p. 175. ISBN: 0412246201.

Smith, B. J. et al. (2001). 'Analysis of inhibitor binding in influenza virus neuraminidase.' In: *Protein Science* vol. 10, no. 4, pp. 689–696.

Smith, G. J. D. et al. (2009). 'Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic'. In: *Nature* vol. 459, no. 7250, pp. 1122–1125.

Sobel Leonard, A. et al. (2016). 'Deep Sequencing of Influenza A Virus from a Human Challenge Study Reveals a Selective Bottleneck and Only Limited Intrahost Genetic Diversification.' In: *Journal of Virology* vol. 90, no. 24, pp. 11247–11258.

Sobel Leonard, A. et al. (2017a). 'The effective rate of influenza reassortment is limited during human infection'. In: *PLOS Pathogens* vol. 13, no. 2, e1006203.

Sobel Leonard, A. et al. (2017b). 'Transmission Bottleneck Size Estimation from Pathogen Deep-Sequencing Data, with an Application to Human Influenza A Virus'. In: *Journal of Virology* vol. 91, no. 14, e00171–17–19.

Stack, J. C. et al. (2013). 'Inferring the inter-host transmission of influenza A virus using patterns of intra-host genetic variation.' In: *Proceedings of the Royal Society B* vol. 280, no. 1750.

Steel, J. et al. (2009). 'Transmission of influenza virus in a mammalian host is increased by PB2 amino acids 627K or 627E/701N.' In: *PLOS Pathogens* vol. 5, no. 1, e1000252.

Stray, S. J. and G. M. Air (2001). 'Apoptosis by influenza viruses correlates with efficiency of viral mRNA synthesis.' In: *Virus Research* vol. 77, no. 1, pp. 3–17.

Strelkowa, N. and M. Lässig (2012). 'Clonal interference in the evolution of influenza.' In: *Genetics* vol. 192, no. 2, pp. 671–682.

Sutton, T. C. et al. (2014). 'Airborne transmission of highly pathogenic H7N1 influenza virus in ferrets.' In: *Journal of Virology* vol. 88, no. 12, pp. 6623–6635.

Tao, H., J. Steel and A. C. Lowen (2014). 'Intrahost Dynamics of Influenza Virus Reassortment'. In: *Journal of Virology* vol. 88, no. 13, pp. 7485–7492.

Tao, H. et al. (2015). 'Influenza A Virus Coinfection through Transmission Can Support High Levels of Reassortment.' In: *Journal of Virology* vol. 89, no. 16, pp. 8453–61.

Tataru, P. et al. (2017). 'Statistical Inference in the Wright-Fisher Model Using Allele Frequency Data.' In: *Systematic Biology* vol. 66, no. 1, e30–e46.

Terhorst, J., C. Schlötterer and Y. S. Song (2015). 'Multi-locus Analysis of Genomic Time Series Data from Experimental Evolution'. In: *PLOS Genetics* vol. 11, no. 4, e1005069–29.

The Academy of Medical Sciences (2005). *Microbial Challenge Studies of Human Volunteers: A guidance document from the Academy of Medical Sci-*

ences. Accessed: 28-12-2018. URL: https://acmedsci.ac.uk/viewFile/publicationDownloads/1127728424.pdf.

The Bill & Melinda Gates Foundation (2018). *Ending the Pandemic Threat: A Grand Challenge for Universal Influenza Vaccine Development*. Accessed: 08-01-2019. URL: https://gcgh.grandchallenges.org/challenge/ending-pandemic-threat-grand-challenge-universal-influenza-vaccine-development.

The Wellcome Trust (2018a). *Human Infection Studies for Vaccine Development*. Accessed: 30-12-2018. URL: https://wellcome.ac.uk/funding/human-infection-studies-vaccine-development.

— (2018b). *Priority area - Vaccines: a world equipped to combat infectious disease*. Accessed: 30-12-2018. URL: https://wellcome.ac.uk/what-we-do/our-work/vaccines.

Thomas, S. J. (2013). 'Dengue human infection model'. In: *Human Vaccines & Immunotherapeutics* vol. 9, no. 7, pp. 1587–1590.

Tran, T. D., J. Hofrichter and J. Jost (2014). 'The evolution of moment generating functions for the Wright-Fisher model of population genetics.' In: *Mathematical Biosciences* vol. 256, pp. 10–17.

Tsai, Y., F. Zhou and I. K. Kim (2014). 'The burden of influenza-like illness in the US workforce.' In: *Occupational Medicine* vol. 64, no. 5, pp. 341–7.

UK Cabinet Office (2017). *National Risk Register Of Civil Emergencies*. Accessed: 08-01-2019. URL: https://www.gov.uk/government/publications/national-risk-register-of-civil-emergencies-2017-edition.

Varble, A. et al. (2014). 'Influenza A virus transmission bottlenecks are defined by infection route and recipient host.' In: *Cell Host & Microbe* vol. 16, no. 5, pp. 691–700.

Varghese, V. et al. (2010). 'Nucleic Acid Template and the Risk of a PCR-Induced HIV-1 Drug Resistance Mutation'. In: *PLOS ONE* vol. 5, no. 6, e10992–6.

Visher, E. et al. (2016). 'The Mutational Robustness of Influenza A Virus'. In: *PLOS Pathogens* vol. 12, no. 8, e1005856–25.

Visser, J. A. G. M. de and J. Krug (2014). 'Empirical fitness landscapes and the predictability of evolution.' In: *Nature Reviews Genetics* vol. 15, no. 7, pp. 480–490.

Wahl, H. R., G. B. White and H. W. Lyall (1919). 'Some Experiments on the Transmission of Influenza'. In: *Journal of Infectious Diseases* vol. 25, no. 5, pp. 419–426.

Wang, C. et al. (2007). 'Characterization of mutation spectra with ultra-deep pyrosequencing: Application to HIV-1 drug resistance'. In: *Genome Research* vol. 17, no. 8, pp. 1195–1201.

Watanabe, T. et al. (2014). 'Circulating avian influenza viruses closely related to the 1918 virus have pandemic potential.' In: *Cell Host & Microbe* vol. 15, no. 6, pp. 692–705.

Webster, R. G. et al. (1992). 'Evolution and ecology of influenza A viruses.' In: *Microbiological Reviews* vol. 56, no. 1, pp. 152–79.

Wilker, P. R. et al. (2013). 'Selection on haemagglutinin imposes a bottleneck during mammalian transmission of reassortant H5N1 influenza viruses.' In: *Nature Communications* vol. 4, p. 2636.

Wilkinson, T. M. et al. (2012). 'Preexisting influenza-specific CD4+ T cells correlate with disease protection against influenza challenge in humans'. In: *Nature Medicine* vol. 18, no. 2, pp. 274–280.

Woods, C. W. et al. (2013). 'A Host Transcriptional Signature for Presymptomatic Detection of Infection in Humans Exposed to Influenza H1N1 or H3N2'. In: *PLOS ONE* vol. 8, no. 1, e52198.

Worby, C. J., M. Lipsitch and W. P. Hanage (2017). 'Shared Genomic Variants: Identification of Transmission Routes Using Pathogen Deep-Sequence Data'. In: *American Journal of Epidemiology* vol. 186, no. 10, pp. 1209–1216.

World Health Organization (2010). *Pandemic (H1N1) 2009 - update 112*. Accessed: 03-12-2018. URL: https://www.who.int/csr/don/2010_08_06/en/.

— (2018a). *Cumulative number of confirmed human cases for avian influenza A(H5N1) reported to WHO, 2003-2018*. Accessed: 30-12-2018. URL: https://www.who.int/influenza/human_animal_interface/2018_12_13_tableH5N1.pdf.

— (2018b). *Fact sheet: Influenza (Seasonal)*. Accessed: 08-01-2019. URL: https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal).

— (2018c). 'Typhoid vaccines: WHO position paper – March 2018'. In: *Weekly epidemiological record* vol. 93, no. 13, pp. 153–172.

Xue, K. S. and J. D. Bloom (2018). 'Reconciling disparate estimates of viral genetic diversity during human influenza infections'. In: *bioRxiv*. eprint: https://www.biorxiv.org/content/early/2018/07/08/364430.full.pdf.

Xue, K. S. et al. (2017). 'Parallel evolution of influenza across multiple spatiotemporal scales'. In: *eLife* vol. 6, e26875.

Yamanouchi, P., K. Sakakami and S. Iwashima (1919). 'The infecting agent in influenza: An experimental research'. In: *The Lancet* vol. 193, no. 4997, p. 971.

Yang, H. et al. (2016). 'Prevalence, genetics, and transmissibility in ferrets of Eurasian avian-like H1N1 swine influenza viruses.' In: *Proceedings of the National Academy of Sciences* vol. 113, no. 2, pp. 392–397.

Yen, H.-L. et al. (2011). 'Hemagglutinin–neuraminidase balance confers respiratory-droplet transmissibility of the pandemic H1N1 influenza virus in ferrets'. In: *Proceedings of the National Academy of Sciences of the United States of America* vol. 108, no. 34, pp. 14264–14269.

Zaas, A. K. et al. (2009). 'Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans.' In: *Cell Host & Microbe* vol. 6, no. 3, pp. 207–17.

Zaas, A. K. et al. (2013). 'A host-based RT-PCR gene expression signature to identify acute respiratory viral infection.' In: *Science Translational Medicine* vol. 5, no. 203, 203ra126.

Zanini, F. et al. (2015). 'Population genomics of intrapatient HIV-1 evolution'. In: *eLife* vol. 4, e11282.

Zanini, F. et al. (2017). 'Error rates, PCR recombination, and sampling depth in HIV-1 whole genome deep sequencing.' In: *Virus Research* vol. 239, pp. 106–114.

Zaraket, H. et al. (2015). 'Mammalian adaptation of influenza A(H7N9) virus is limited by a narrow genetic bottleneck.' In: *Nature Communications* vol. 6, p. 6553.

Zürcher, T. et al. (2006). 'Mutations conferring zanamivir resistance in human influenza virus N2 neuraminidases compromise virus fitness and are not stably maintained in vitro.' In: *The Journal of Antimicrobial Chemotherapy* vol. 58, no. 4, pp. 723–732.

Zwart, M. P., J.-A. Daròs and S. F. Elena (2011). 'One Is Enough: In Vivo Effective Population Size Is Dose-Dependent for a Plant RNA Virus'. In: *PLOS Pathogens* vol. 7, no. 7, e1002122–12.

Zwart, M. P. and S. F. Elena (2015). 'Matters of Size: Genetic Bottlenecks in Virus Infection and Their Potential Impact on Evolution'. In: *Annual Review of Virology* vol. 2, no. 1, pp. 161–179.

Appendix A

Discrete Compound Solution

A.1 Introduction

Given a resulting mean vector, $\boldsymbol{\mu}$, and a covariance matrix, Σ, arising from a series of compounding events, we may use these moments for computing likelihoods under a multivariate normal scheme. Alternatively, under certain circumstances, it may be possible to revert to a discretised solution by associating the resulting moments with the mean and variance of a Dirichlet-multinomial distribution. Here I first present a derivation of the parameter $\boldsymbol{\alpha}$ describing the Dirichlet-multinomial compound solution, noting potential limitations with regards to discretisation. Next, I perform a derivation of the shape parameters α and β of the beta-binomial distribution, exploring here in detail discretisation constraints in the one-dimensional setting.

A.2 Discrete Solution - Multi-Dimensional Setting

The mean and variance of a Dirichlet-multinomial distribution are related to the parameter $\boldsymbol{\alpha}$ as follows:

$$\boldsymbol{\mu} = n\frac{\boldsymbol{\alpha}}{\alpha_0} \tag{A.1}$$

and

$$\Sigma = n\left(\text{Diag}\left(\frac{\boldsymbol{\alpha}}{\alpha_0}\right) - \left(\frac{\boldsymbol{\alpha}}{\alpha_0}\right)\left(\frac{\boldsymbol{\alpha}}{\alpha_0}\right)^{\dagger}\right)\left(\frac{n+\alpha_0}{1+\alpha_0}\right) \tag{A.2}$$

where $\alpha_0 = \sum_i^k \alpha_i$.

From the equation for the mean we have that $\boldsymbol{\alpha} = \frac{\alpha_0 \boldsymbol{\mu}}{n}$, which we may substitute into the equation for the variance:

$$\begin{aligned} \Sigma &= n\left(\text{Diag}\left(\frac{\boldsymbol{\mu}}{n}\right) - \left(\frac{\boldsymbol{\mu}}{n}\right)\left(\frac{\boldsymbol{\mu}}{n}\right)^\dagger\right)\left(\frac{n+\alpha_0}{1+\alpha_0}\right) \\ &= \left(\text{Diag}\left(\boldsymbol{\mu}\right) - \frac{1}{n}\boldsymbol{\mu}\boldsymbol{\mu}^\dagger\right)\left(\frac{n+\alpha_0}{1+\alpha_0}\right) \end{aligned} \tag{A.3}$$

Now defining $\tilde{M}(n, \boldsymbol{\mu}) = \left(\text{Diag}\left(\boldsymbol{\mu}\right) - \frac{1}{n}\boldsymbol{\mu}\boldsymbol{\mu}^\dagger\right)$, the matrix $\tilde{M}$ is invertible if $\text{Det}(\tilde{M}) \neq 0$. Given a k-dimensional problem, i.e. $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_k\}$, then

$$\text{Det}(\tilde{M}) = \left(\prod_{i=1}^k \mu_i\right)\frac{1}{n^k}\left[-\sum_{i=1}^k \mu_i n^{k-1} + n^k\right] \tag{A.4}$$

so $\text{Det}(\tilde{M}) = 0$ is true only if $\sum_{i=1}^k \mu_i = n$. Now, for k dimensions, this is always true.

Given that the mean can be described fully by $k-1$ dimensions, i.e. the k^{th} dimension being $\mu_k = n - \sum_{i=1}^{k-1} \mu_i$, we can reduce the dimensionality of the problem and as a result the determinant will be non-zero for all cases. As such, $\tilde{M}$ is invertible.

Having found $\tilde{M}^{-1}$, we can rewrite the equation for the covariance matrix as follows:

$$\begin{aligned} \Sigma &= \tilde{M}(n, \boldsymbol{\mu})\left(\frac{n+\alpha_0}{1+\alpha_0}\right) \\ \tilde{M}^{-1}(n, \boldsymbol{\mu})\Sigma &= \left(\frac{n+\alpha_0}{1+\alpha_0}\right)\mathbb{I}^{k-1} \end{aligned} \tag{A.5}$$

where $\mathbb{I}^{k-1}$ is the $k-1$ dimensional identity matrix.

Defining $Q = \tilde{M}^{-1}(n, \boldsymbol{\mu})\Sigma$, then Q must be diagonal in some vector $\boldsymbol{q} = \{q_i\}$. In fact, all the q_i must be identical, which introduces constraints on μ and Σ, these constraints arising from the Dirichlet-multinomial being a more rigid distribution than e.g. the multivariate normal distribution. In turn, it is only possible to generate a discretised solution for a specific category of μ and Σ, this aspect being explored further in the one-dimensional setting below.

As such, with out loss of generality, we may compute α_0 from $Q_{1,1}$:

$$\begin{aligned} Q_{1,1} = q_1 &= \left(\frac{n+\alpha_0}{1+\alpha_0}\right) \\ q_1(1+\alpha_0) &= n+\alpha_0 \\ \alpha_0(1-q_1) &= q_1 - n \\ \alpha_0 &= \frac{n-q_1}{q_1-1} \end{aligned} \tag{A.6}$$

And substituting into the equation for the mean (Equation A.1) yields

$$\boldsymbol{\alpha} = \frac{n-q_1}{n(q_1-1)}\boldsymbol{\mu} \tag{A.7}$$

A.3 Discrete Solution - One-Dimensional Setting

I here attempt to derive and analyse a discrete solution in one dimension. In this case we have a mean μ and a variance σ^2 and aim to derive parameters α and β describing a beta-binomial distribution with a fixed n.

A.3.1 Derivation of α and β

For the beta-binomial distribution we know that

$$\mu = \frac{n\alpha}{\alpha+\beta} \tag{A.8}$$

and

$$\sigma^2 = \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{A.9}$$

Solving the first equation for β gives

$$\beta = \frac{\alpha(n-\mu)}{\mu} \tag{A.10}$$

Substituting this into the equation for the variance and simplifying gives

$$\begin{aligned}
\sigma^2 &= \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)} \\
&= \frac{n\alpha\frac{\alpha(n-\mu)}{\mu}(\alpha+\frac{\alpha(n-\mu)}{\mu}+n)}{(\alpha+\frac{\alpha(n-\mu)}{\mu})^2(\alpha+\frac{\alpha(n-\mu)}{\mu}+1)} \\
&= \frac{n\alpha^2\frac{(n-\mu)}{\mu}(\frac{\alpha\mu+\alpha(n-\mu)+\mu n}{\mu})}{(\frac{\alpha\mu+\alpha(n-\mu)}{\mu})^2(\frac{\alpha\mu+\alpha(n-\mu)+\mu}{\mu})} \\
&= \frac{n\alpha^2\frac{(n-\mu)}{\mu}\frac{n(\alpha+\mu)}{\mu}}{(\frac{\alpha n}{\mu})^2(\frac{\alpha n+\mu}{\mu})} \\
&= \frac{n^2\alpha^2\frac{(n-\mu)(\alpha+\mu)}{\mu^2}}{(\alpha n)^2(\frac{\alpha n+\mu}{\mu^3})} \\
&= \frac{(n-\mu)(\alpha+\mu)}{(\frac{\alpha n+\mu}{\mu})} \\
&= \frac{\mu(n-\mu)(\alpha+\mu)}{\alpha n+\mu}
\end{aligned} \tag{A.11}$$

We may rearrange this for α

$$\begin{aligned}
\alpha n+\mu &= \frac{\mu(n-\mu)(\alpha+\mu)}{\sigma^2} \\
\alpha n+\mu &= \frac{\mu(n\alpha-\mu\alpha+n\mu-\mu^2)}{\sigma^2} \\
\alpha n &= \frac{\mu(n\alpha-\mu\alpha)}{\sigma^2}\frac{\mu(n\mu-\mu^2)}{\sigma^2}-\mu \\
\alpha n+\frac{\mu(\mu\alpha-n\alpha)}{\sigma^2} &= \frac{\mu(n\mu-\mu^2)}{\sigma^2}-\mu \\
\frac{\alpha n\sigma^2+\alpha\mu(\mu-n)}{\sigma^2} &= \frac{\mu(n\mu-\mu^2)-\mu\sigma^2}{\sigma^2} \\
\alpha(n\sigma^2+\mu(\mu-n)) &= \mu(n\mu-\mu^2)-\mu\sigma^2 \\
\alpha &= \frac{\mu(n\mu-\mu^2)-\mu\sigma^2}{n\sigma^2+\mu(\mu-n)} \\
\alpha &= \frac{\mu^2(n-\mu)-\mu\sigma^2}{n\sigma^2+\mu(\mu-n)}
\end{aligned} \tag{A.12}$$

or more expressively,

$$\alpha = \frac{-\mu^3+\mu^2 n-\mu\sigma^2}{\mu^2-\mu n+n\sigma^2} \tag{A.13}$$

which in turns gives a β of

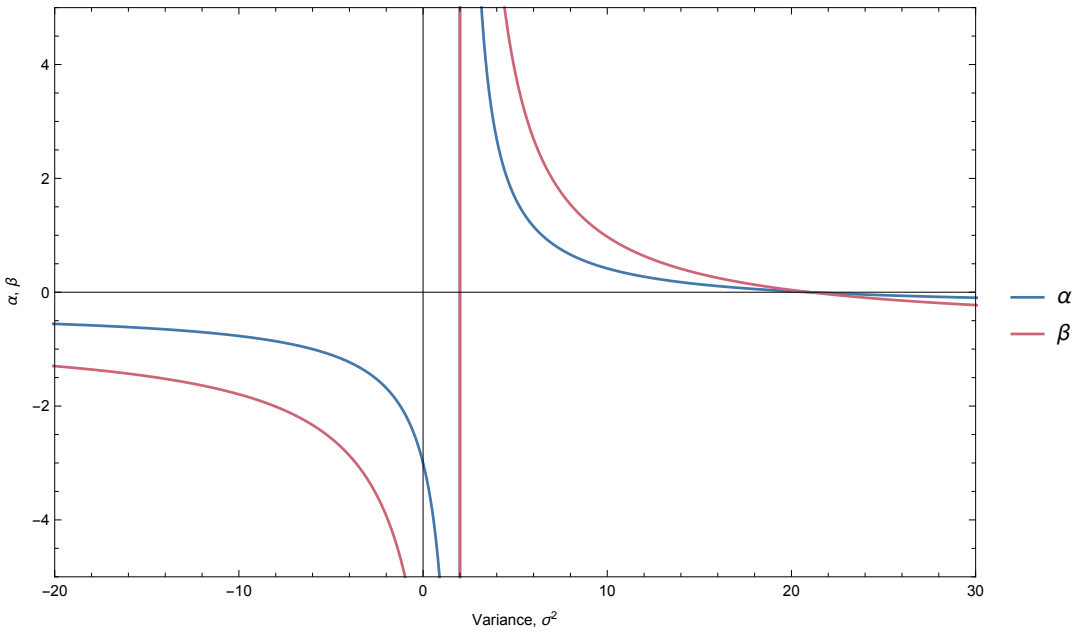


Figure A.1. Plots of α and β as a function of the compound variance σ^2 under a beta-binomial parameterisation, here shown with $\mu = 3$ and $n = 10$. A vertical asymptote is observed at $\sigma^2 = 2.1$, with horizontal asymptotes occurring for $\alpha, \beta < 0$. Intersection with the x-axis is observed at $\sigma^2 = 21$ for both α and β.

$$\begin{aligned} \beta &= \frac{\alpha(n-\mu)}{\mu} \\ \beta &= \frac{-\mu^3 + \mu^2 n - \mu\sigma^2}{\mu^2 - \mu n + n\sigma^2}\frac{(n-\mu)}{\mu} \\ \beta &= \frac{(\mu - n)(\mu^2 - \mu n + \sigma^2)}{\mu^2 - \mu n + n\sigma^2} \end{aligned} \tag{A.14}$$

A.3.2 Limitations on the Variance

We immediately note that α and β have the same denominator. This denominator gives rise to a vertical asymptote in α and β at $\sigma^2 = \frac{\mu n - \mu^2}{n}$. As such, this equation sets a limit on how small the variance can be for the parameterisation of α and β to be sensible (the regime wherein they take finite, positive values). Therefore we cannot necessarily convert any mean and variance into parameters for a beta-binomial; some restrictions apply. A graphical depiction of the vertical asymptote is shown in Figure A.1 for $\mu = 3$ and $n = 10$. We here note that the asymptote is at $\frac{\mu n - \mu^2}{n} = 2.1$ in this specific case.

Additionally we note that α has a horizontal asymptote at $\frac{-\mu}{n}$ whilst β has an asymptote at $\frac{-\mu(n-\mu)}{n\mu} = \frac{-(n-\mu)}{n}$. We note that both asymptotes lie below the

x-axis provided that $\mu < n$, which should in general be true for the compound solutions we consider. The horizontal asymptotes for $\mu = 3$ and $n = 10$ can be seen in Figure A.1. We may also deduce the points at which the graphs for α and β cross the x-axis. For α, this occurs at

$$\begin{aligned} \alpha &= \frac{-\mu^3 + \mu^2 n - \mu\sigma^2}{\mu^2 - \mu n + n\sigma^2} = 0 \\ 0 &= -\mu^3 + \mu^2 n - \mu\sigma^2 \\ \sigma^2 &= \mu n - \mu^2 = \mu(n - \mu) \end{aligned} \tag{A.15}$$

whilst for β it occurs at

$$\begin{aligned} \beta &= \frac{(\mu - n)(\mu^2 - \mu n + \sigma^2)}{\mu^2 - \mu n + n\sigma^2} = 0 \\ 0 &= (\mu - n)(\mu^2 - \mu n + \sigma^2) \\ 0 &= \mu^3 - \mu^2 n + \mu\sigma^2 - \mu^2 n + \mu n^2 - n\sigma^2 \\ 0 &= \mu^3 - 2\mu^2 n + \mu n^2 - (n - \mu)\sigma^2 \\ \sigma^2 &= \frac{\mu^3 - 2\mu^2 n + \mu n^2}{n - \mu} \\ \sigma^2 &= \frac{\mu(\mu^2 - 2\mu n + n^2)}{n - \mu} \\ \sigma^2 &= \frac{\mu(n - \mu)^2}{n - \mu} \\ \sigma^2 &= \mu(n - \mu) \end{aligned} \tag{A.16}$$

i.e. the same intersect as α. For $\mu = 3$ and $n = 10$ the intersect is at $\sigma^2 = 21$ as can be seen in Figure A.1.

As such, we have defined lower and upper bounds of the variance which allows for positive α and β, namely $\sigma^2 \in \left(\frac{\mu n - \mu^2}{n}, \mu(n - \mu)\right)$. Variance-mean configurations outside this range represents compound solutions that cannot be represented using a discrete parameterisation.

Appendix B

Examination of One-Dimensional Discrete Solution

B.1 Introduction

In Appendix A I derived a one-dimensional beta-binomial framework for evaluating compound solutions defined by a mean μ and variance σ^2. In this appendix I employ the solution in a simple transmission setup, comparing its performance to a Gaussian approach.

B.2 Transmission Model and Compound Solution

We employ a transmission model similar to that of Figure 3.2, the only difference being that we now consider a one-dimensional problem. To this extend the likelihood framework of Chapter 3 is unchanged, differences occurring only in the compound solutions. For a purely neutral setup, the one-dimensional solution has mean

$$\mathrm{E}[x^A] = N^A \mu^B \tag{B.1}$$

and variance

$$\mathrm{var}(\boldsymbol{x}^A) = N^A \left(\alpha + (N^A - \alpha)\gamma\right) \mu^B (1 - \mu^B) + N^A \left(N^A - \alpha\right) \delta \left(\sigma^B\right)^2 \tag{B.2}$$

where $\gamma = \left(\frac{N^T + N^G - 1}{N^T N^G}\right)$, $\delta = \frac{N^T N^G - N^T - N^G + 1}{N^T N^G}$ and $\alpha = \frac{N^A + C}{1 + C}$ as previously.

These expressions may be straightforwardly derived by following an approach similar to that described in Appendix C. Alternatively, the expressions are readily obtained simply by comparing with the final two equations of Appendix C.

Under selection for transmission the compound solution has mean

$$\mathrm{E}[x^A] = N^A S^T\left(\mu^B\right) \tag{B.3}$$

and variance

$$\begin{aligned}\mathrm{var}(\boldsymbol{x}^A) = N^A\left(\alpha + (N^A-\alpha)\gamma\right) S^T(\mu^B)\left(1 - S^T(\mu^B)\right) \\ + N^A\left(N^A-\alpha\right)\delta\left(\left(S^T\right)'\Big|_{\mu^B}\right)^2\left(\sigma^B\right)^2\end{aligned} \tag{B.4}$$

with γ, δ and α as above and with prime ($'$) denoting differentiation with respect to q^B in

$$\left(S^T(q^B)\right) = \frac{w^T q^B}{w^T q^B + (1-q^B)} \tag{B.5}$$

where w^T is the fitness effect of harbouring the haplotype corresponding to the frequency q^B. Fitnesses are evaluated in a manner similar to that outlined in Section 3.2.4.1.

These expressions may straightforwardly be derived by following an approach similar to that described in Chapter 3, or simply by comparison with Equations 3.29 and 3.30.

B.3 Comparison of Beta-Binomial and Gaussian Solutions

In the following we consider a number of different transmission scenarios and compare the beta-binomial solution (Appendix A) to the Gaussian solution.

B.3.1 Neutral Transmission

First we consider three neutral transmission events with varying parameters of transmission. In the first instance we consider the transmission of a population with before frequency $q^B = 0.3$ under a bottleneck of size $N^T = 20$. The effective growth size and sampling depth are set to $N^G = N^A = 100$ whilst the extend of

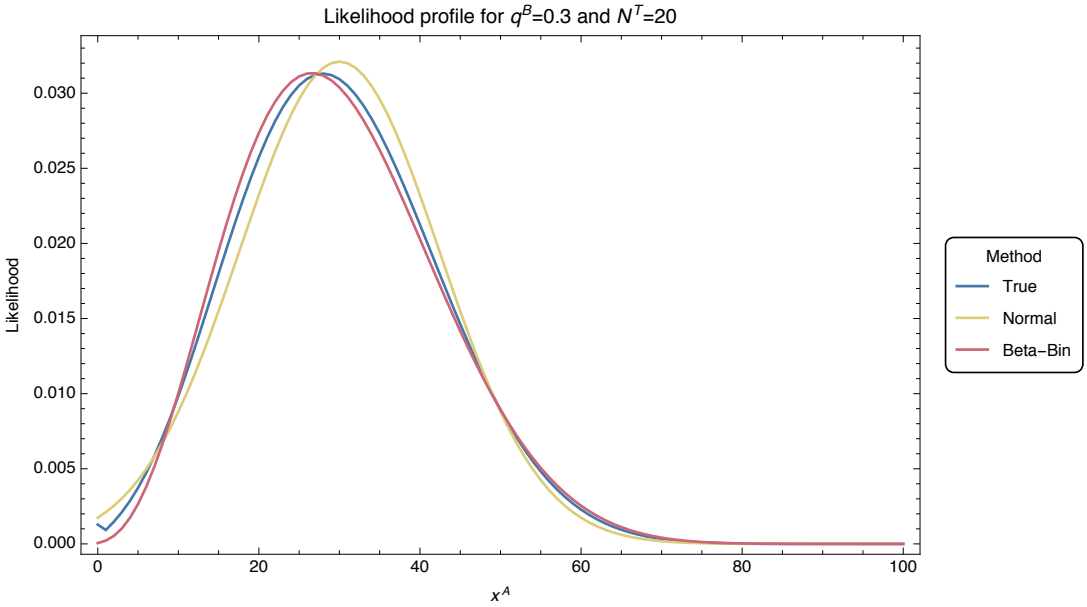


Figure B.1. Likelihood profile for the true distribution, the normal distribution and the beta-binomial distribution for a transmission with $q^B = 0.3$ and $N^T = 20$. The sampling depths are $N^G = N^A = 100$ and the extent of noise is $C = 200$. For the compound solution we used $C^{\text{inf}} = C$, $\mu^B = q^B$ and $\left(\sigma^B\right)^2 = 10^{-6}$.

noise is $C = 200$. For the compound solution we employ an inferred C-value of $C^{\text{inf}} = C$, a mean of $\mu^B = q^B$ and a variance of $\left(\sigma^B\right)^2 = 10^{-6}$, i.e. we are very certain in our inference of the before population. The likelihood profiles for the true distribution, the normal distribution and the beta-binomial distribution are shown in Figure B.1. The true and beta-binomial distributions have been plotted as continuous curves in order to improve readability and facilitate comparison. We here observe that the three distributions are very similar, but that the beta-binomial approach is marginally better at approximating the true distribution than the normal method is. This is because the beta-binomial distribution takes into account skewness, whereas the normal distribution is entirely symmetric.

In Figure B.2 we consider an identical setup, except that the true and inferred before frequencies are changed to $q^B = \mu^B = 0.1$. As expected, the three distributions are now shifted towards $x^A = 0$. The true distribution displays a peak around $x^A = 6$ and a spike at $x^A = 0$. This may be understood as an increased probability of extinction when the before population lies near the boundary. The normal distribution does not account for this, having a peak at $x^A = 10$. The beta-binomial distribution attempts to account for this behaviour, having a peak at around $x^A = 3$. Regardless, both distribution perform poorly near the boundaries.

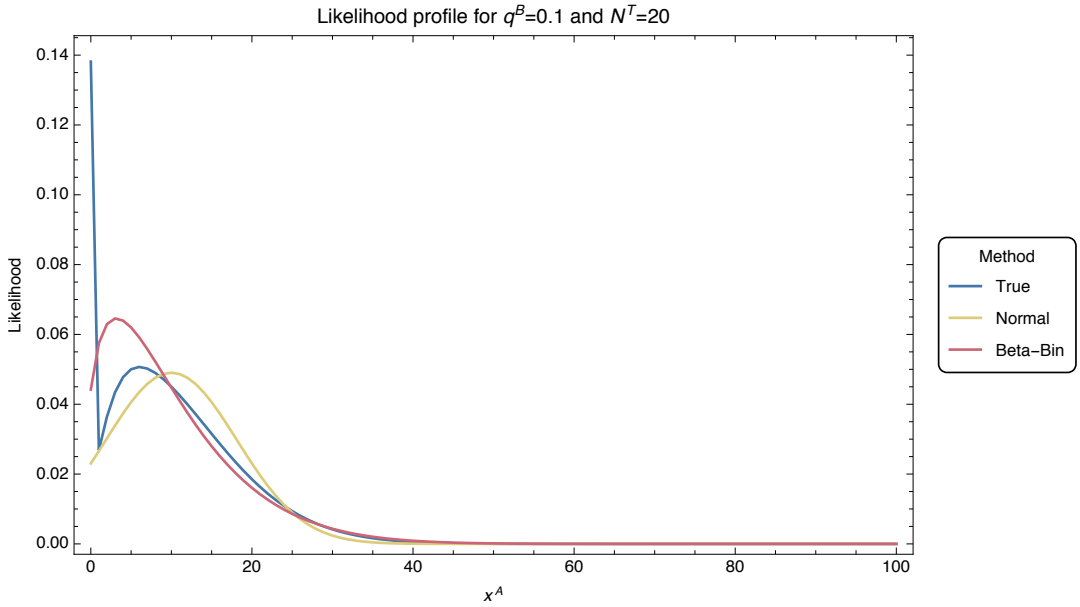


Figure B.2. Likelihood profile for the true distribution, the normal distribution and the beta-binomial distribution for a transmission with $q^B = 0.1$ and $N^T = 20$. The sampling depths are $N^G = N^A = 100$ and the extent of noise is $C = 200$. For the compound solution we used $C^{\text{inf}} = C$, $\mu^B = q^B$ and $\left(\sigma^B\right)^2 = 10^{-6}$.

In Figure B.3 we consider a setup identical to that of Figure B.1, but here we alter the uncertainty in our inference of μ^B, increasing the variance to $\left(\sigma^B\right)^2 = 0.01$. Here, the normal distribution becomes distinctly wider, but keeps the mean fixed. Conversely, the beta-binomial distribution is shifted considerably to the left. This may be explained by the beta-binomial distribution translating variance into skewness. This suggests that the beta-binomial method only works under the assumption of complete knowledge of μ^B.

B.3.2 Selection for Transmission

Finally we examine a case of a transmission event governed by selection for transmission. Here we consider the transmission of a population with before frequency $q^B = 0.3$ under a bottleneck of size $N^T = 20$. We here employ selective pressure of magnitude $\sigma^T = 2$, i.e. a shift in frequency towards $x^A = N^A$. The effective growth size and sampling depth are set to $N^G = N^A = 100$ whilst the extend of noise is $C = 200$. For the compound solution we employ an inferred C-value of $C^{\text{inf}} = C$, a mean of $\mu^B = q^B$, a variance of $\left(\sigma^B\right)^2 = 10^{-6}$ and inferred selection of $\sigma^{T,\text{inf}} = \sigma^T$. The likelihood profiles are shown in Figure B.4. As expected, the three distributions are shifted towards $x^A = 100$. We here notice

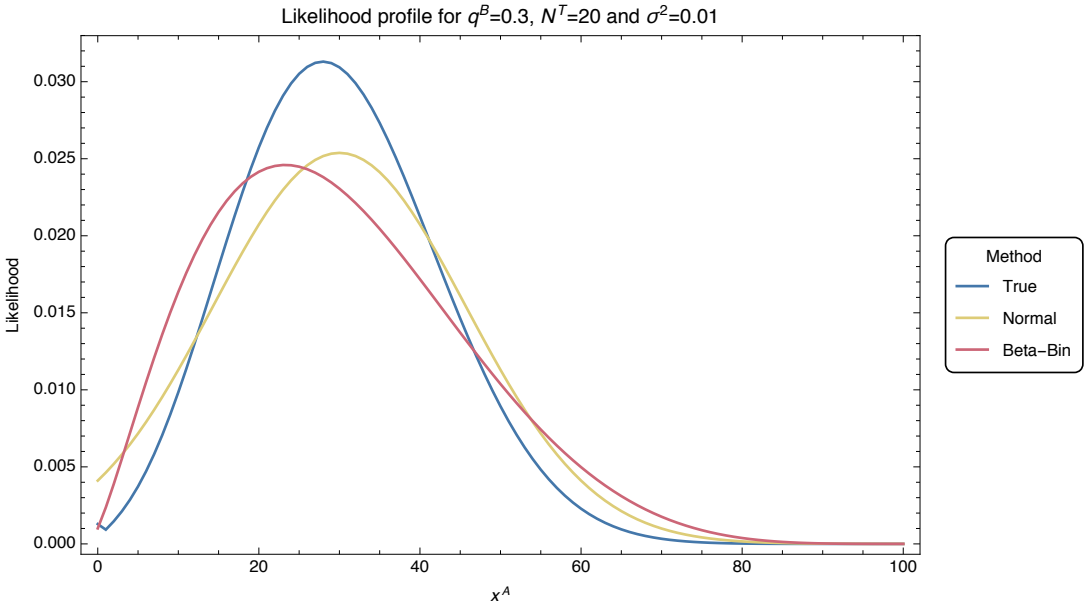


Figure B.3. Likelihood profile for the true distribution, the normal distribution and the beta-binomial distribution for a transmission with $q^B = 0.3$ and $N^T = 20$. The sampling depths are $N^G = N^A = 100$ and the extent of noise is $C = 200$. For the compound solution we used $C^{\text{inf}} = C$, $\mu^B = q^B$ and $\left(\sigma^B\right)^2 = 0.01$.

that the beta-binomial method slightly outperforms the Gaussian approach as it is able to account for skewness.

B.4 Summary

In conclusion, I observe that the beta-binomial method has some distinct advantages over the Gaussian approach, primarily due to its ability to correctly account for skewness. However, a major flaw is the inability of the beta-binomial method to account for uncertainty in the inference of μ^B. Furthermore, as I have shown in Appendix A, the beta-binomial discretisation framework is only valid for a certain regime of μ and σ (not to be confused with μ^B and σ^B). As a result, I have not explored discrete approaches further.

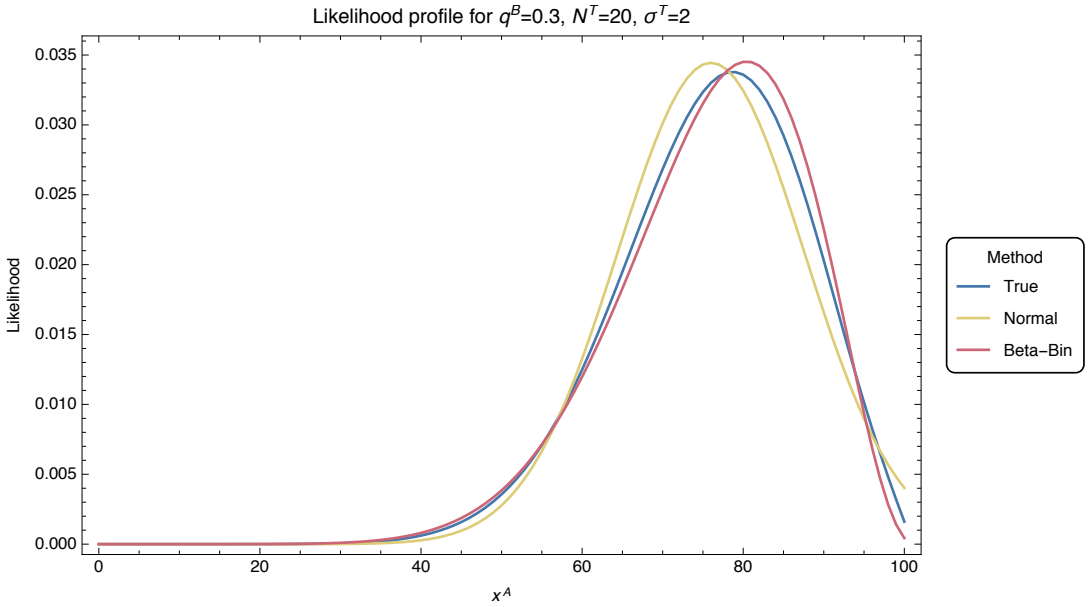


Figure B.4. Likelihood profile for the true distribution, the normal distribution and the beta-binomial distribution for a transmission with $q^B = 0.3$, $N^T = 20$ and $\sigma^T = 2$. The sampling depths are $N^G = N^A = 100$ and the extent of noise is $C = 200$. For the compound solution we used $C^{\text{inf}} = C$, $\mu^B = q^B$, $\left(\sigma^B\right)^2 = 10^-6$ and $\sigma^{T,\text{inf}} = \sigma^T$.

Appendix C

Compound Solution for Basic Model Under Neutrality

C.1 Introduction

The resulting mean and variance of the likelihood expression defined in Equation 3.16 may be derived in the absence of selection. This may be regarded as a special case of Equations 3.29 and 3.30. Under neutrality the founder populations has mean

$$\mathrm{E}[\boldsymbol{q}^F|\boldsymbol{q}^B] = \boldsymbol{q}^B \tag{C.1}$$

and variance

$$\mathrm{var}[\boldsymbol{q}^F|\boldsymbol{q}^B] = \frac{1}{N^T} M(\boldsymbol{q}^B) \tag{C.2}$$

with the remaining conditional moments as given in Chapter 3.

C.2 Derivation

The marginalisation over $\boldsymbol{q}^B$ results in a mean of

$$\mathrm{E}[\boldsymbol{q}^F] = \mathrm{E}[\mathrm{E}[\boldsymbol{q}^F|\boldsymbol{q}^B]] = \mathrm{E}[\boldsymbol{q}^B] = \boldsymbol{\mu}^B \tag{C.3}$$

whilst the law of total variance yields

$$\begin{aligned}\text{var}(\boldsymbol{q}^F) &= \text{E}[\text{var}[\boldsymbol{q}^F|\boldsymbol{q}^B]] + \text{var}[\text{E}[\boldsymbol{q}^F|\boldsymbol{q}^B]] \\ &= \text{E}\left[\frac{1}{N^T}M(\boldsymbol{q}^B)\right] + \text{var}\left[\boldsymbol{q}^B\right] \\ &= \frac{1}{N^T}M(\text{E}[\boldsymbol{q}^B]) + \left(1 - \frac{1}{N^T}\right)\text{var}[\boldsymbol{q}^B] \\ &= \frac{1}{N^T}M(\boldsymbol{\mu}^B) + \left(1 - \frac{1}{N^T}\right)\Sigma^B\end{aligned} \tag{C.4}$$

Marginalisation over $\boldsymbol{q}^F$ yields a mean of

$$\text{E}[\boldsymbol{q}^A] = \text{E}[\text{E}[\boldsymbol{q}^A|\boldsymbol{q}^F]] = \text{E}[\boldsymbol{q}^F] = \boldsymbol{\mu}^B \tag{C.5}$$

and variance

$$\begin{aligned}\text{var}(\boldsymbol{q}^A) &= \text{E}[\text{var}[\boldsymbol{q}^A|\boldsymbol{q}^F]] + \text{var}[\text{E}[\boldsymbol{q}^A|\boldsymbol{q}^F]] \\ &= \text{E}\left[\frac{1}{N^G}\left(\text{Diag}(\boldsymbol{q}^F) - \boldsymbol{q}^F(\boldsymbol{q}^F)^\dagger\right)\right] + \text{var}[\boldsymbol{q}^F] \\ &= \frac{1}{N^G}\left(\text{Diag}(\text{E}[\boldsymbol{q}^F]) - \text{E}[\boldsymbol{q}^F]E[\boldsymbol{q}^F]^\dagger\right) + \left(1 - \frac{1}{N^G}\right)\text{var}[\boldsymbol{q}^F] \\ &= \frac{1}{N^G}M\left(\boldsymbol{\mu}^B\right) + \left(1 - \frac{1}{N^G}\right)\left(\frac{1}{N^T}M(\boldsymbol{\mu}^B) + \left(1 - \frac{1}{N^T}\right)\Sigma^B\right) \\ &= \frac{N^T + N^G - 1}{N^T N^G}M\left(\boldsymbol{\mu}^B\right) + \frac{N^T N^G - N^T - N^T + 1}{N^T N^G}\Sigma^B \\ &\equiv \gamma M\left(\boldsymbol{\mu}^B\right) + \delta\Sigma^B\end{aligned} \tag{C.6}$$

where in the last step we defined $\gamma = \left(\frac{N^T+N^G-1}{N^T N^G}\right)$ and $\delta = \frac{N^T N^G - N^T - N^G + 1}{N^T N^G}$.

Treating the integral over $\boldsymbol{q}^A$ in a similar manner, we obtain by the law of total expectation

$$\text{E}[\boldsymbol{x}^A] = \text{E}[\text{E}[\boldsymbol{x}^A|\boldsymbol{q}^A]] = \text{E}[N^A\boldsymbol{q}^A] = N^A\text{E}[\boldsymbol{q}^A] = N^A\boldsymbol{\mu}^B \tag{C.7}$$

Analogously, the law of total variance yields

$$\begin{aligned}
\text{var}(\boldsymbol{x}^A) &= \text{E}[\text{var}[\boldsymbol{x}^A|\boldsymbol{q}^A]] + \text{var}[\text{E}[\boldsymbol{x}^A|\boldsymbol{q}^A]] \\
&= \text{E}\left[\alpha N^A M\left(\boldsymbol{q}^A\right)\right] + \text{var}[N^A\boldsymbol{q}^A] \\
&= \alpha N^A\left(\text{Diag}(\text{E}\left[\boldsymbol{q}^A\right] - \text{E}\left[\boldsymbol{q}^A\right]\text{E}\left[\boldsymbol{q}^A\right]^\dagger\right) + N^A\left(N^A-\alpha\right)\text{var}[\boldsymbol{q}^A] \\
&= \alpha N^A M(\boldsymbol{\mu}^B) + N^A\left(N^A-\alpha\right)\left(\gamma M\left(\boldsymbol{\mu}^B\right) + \delta\Sigma^B\right) \\
&= N^A\left(\alpha + (N^A-\alpha)\gamma\right)M(\boldsymbol{\mu}^B) + N^A\left(N^A-\alpha\right)\delta\Sigma^B
\end{aligned} \tag{C.8}$$

These equations replace Equations 3.29 and 3.30 in the absence of selection.

Appendix D

Derivation of Compound Distributions for N-Step Drift Process

D.1 Introduction

This considers the derivation of compound distributions for a general N-step drift process in the presence of selection for A) within-host adaptation and B) both transmission and within-host evolution. This appendix follows on from Section 4.2.3.

D.2 Selection for Within-Host Adaptation

Deriving compound distributions under selection for within-host adaptation is a little more challenging. Two approaches can be taken: A) We assume that selection only acts after all the growth steps, i.e. it acts in the $\boldsymbol{q}^A$ compound. This can straightforwardly be derived by combining results from Section 4.2.2 with the coefficients γ_n and δ_n. B) We assume that selection acts once every 12 hours of growth, i.e. we have interleaving drift and selection processes: $\boldsymbol{q}^F \rightarrow \boldsymbol{q}^{G_1} \rightarrow S^G(\boldsymbol{q}^{G_1}) \rightarrow \boldsymbol{q}^{G_2} \rightarrow S^G(\boldsymbol{q}^{G_2}) \rightarrow \ldots$. In the below we use the latter approach to derive compound solutions in the absence of selection for transmission.

We now outline the conditional distributions required for computing the compound solutions. As previously we have a founder population defined by

$$\boldsymbol{q}^F \sim \mathcal{N}\left(\boldsymbol{q}^B, \tfrac{1}{N^T} M(\boldsymbol{q}^B)\right) \tag{D.1}$$

where $\mathcal{N}$ denotes the multivariate normal distribution.

First we have a round of growth during which no selection applies:

$$\boldsymbol{q}^{G_1} \sim \mathcal{N}\left(\boldsymbol{q}^F, \tfrac{1}{N^{G_1}} M(\boldsymbol{q}^F)\right) = \mathcal{N}\left(\boldsymbol{q}^F, \tfrac{1}{\lambda N^T} M(\boldsymbol{q}^F)\right) \tag{D.2}$$

where we defined $N^{G_1} = \lambda N^T$.

Next we have a growth round during which selection acts (prior to growth):

$$\boldsymbol{q}^{G_2} \sim \mathcal{N}\left(S^G(\boldsymbol{q}^{G_1}), \tfrac{1}{N^{G_2}} M(S^G(\boldsymbol{q}^{G_1}))\right) = \mathcal{N}\left(S^G(\boldsymbol{q}^{G_1}), \tfrac{1}{\lambda^2 N^T} M(S^G(\boldsymbol{q}^{G_1}))\right) \tag{D.3}$$

where S^G is in 12-hour units and $N^{G_2} = \lambda^2 N^T$.

In the same manner, every subsequent growth round will incorporate both growth and selection with selection acting first. In general we have that

$$\begin{aligned} \boldsymbol{q}^{G_n} &\sim \mathcal{N}\left(S^G(\boldsymbol{q}^{G_{n-1}}), \tfrac{1}{N^{G_n}} M(S^G(\boldsymbol{q}^{G_{n-1}}))\right) \\ &= \mathcal{N}\left(S^G(\boldsymbol{q}^{G_{n-1}}), \tfrac{1}{\lambda^n N^T} M(S^G(\boldsymbol{q}^{G_{n-1}}))\right) \end{aligned} \tag{D.4}$$

with $N^{G_n} = \lambda^n N^T$ for $n > 1$.

Assuming N steps in the growth process (e.g. if one step = 12 hours, then $N = 2$ steps would correspond to a 24 hour difference between donor and recipient sampling times), we have

$$\boldsymbol{q}^A \sim \mathcal{N}\left(S^G(\boldsymbol{q}^{G_{N-1}}), \tfrac{1}{N^{G_N}} M(S^G(\boldsymbol{q}^{G_{N-1}}))\right) \tag{D.5}$$

Finally, as always, the post-transmission sampling step is defined by

$$\boldsymbol{x}_i^{A,\mathrm{P}} \sim \mathcal{N}\left(N_i^A T_i S^G(\boldsymbol{q}^A), \alpha_i N_i^A M(T_i S^G(\boldsymbol{q}^A)\right) \tag{D.6}$$

where $\alpha_i = \frac{N_i^A + C}{1+C}$. Note that selection acts prior to sampling here. This ensures that there are equally many drift and selection steps during viral growth.

Turning now to the evaluation of compound distributions, the marginalisation over $\boldsymbol{q}^B$ leads to

$$\mathrm{E}[\boldsymbol{q}^F] = \mathrm{E}[\mathrm{E}[\boldsymbol{q}^F|\boldsymbol{q}^B]] = \mathrm{E}[\boldsymbol{q}^B] = \boldsymbol{\mu}^B \tag{D.7}$$

and,

$$\begin{aligned}
\mathrm{var}(\boldsymbol{q}^F) &= \mathrm{E}[\mathrm{var}[\boldsymbol{q}^F|\boldsymbol{q}^B]] + \mathrm{var}[\mathrm{E}[\boldsymbol{q}^F|\boldsymbol{q}^B]] \\
&= \mathrm{E}\left[\frac{1}{N^T} M(\boldsymbol{q}^B)\right] + \mathrm{var}\left[\boldsymbol{q}^B\right] \\
&= \frac{1}{N^T} M(\mathrm{E}[\boldsymbol{q}^B]) + \left(1 - \frac{1}{N^T}\right) \mathrm{var}[\boldsymbol{q}^B] \\
&= \frac{1}{N^T} M(\boldsymbol{\mu}^B) + \left(1 - \frac{1}{N^T}\right) \Sigma^B \\
&= \gamma_0 M(\boldsymbol{\mu}^B) + \delta_0 \Sigma^B
\end{aligned} \tag{D.8}$$

where in the last step we defined $\gamma_0 = \frac{1}{N^T}$ and $\delta_0 = \left(1 - \frac{1}{N^T}\right)$ (for reasons that will become clear later).

Next, for the $\boldsymbol{q}^F$ integral, the law of total expectation yields

$$\mathrm{E}[\boldsymbol{q}^{G_1}] = \mathrm{E}[\mathrm{E}[\boldsymbol{q}^{G_1}|\boldsymbol{q}^F]] = \mathrm{E}[\boldsymbol{q}^F] = \boldsymbol{\mu}^B \tag{D.9}$$

Next, under the law of total variance,

$$\begin{aligned}
\mathrm{var}(\boldsymbol{q}^{G_1}) &= \mathrm{E}[\mathrm{var}[\boldsymbol{q}^{G_1}|\boldsymbol{q}^F]] + \mathrm{var}[\mathrm{E}[\boldsymbol{q}^{G_1}|\boldsymbol{q}^F]] \\
&= \mathrm{E}\left[\frac{1}{\lambda N^T}\left(\mathrm{Diag}(\boldsymbol{q}^F) - \boldsymbol{q}^F(\boldsymbol{q}^F)^\dagger\right)\right] + \mathrm{var}[\boldsymbol{q}^F] \\
&= \frac{1}{\lambda N^T}\left(\mathrm{Diag}(\mathrm{E}[\boldsymbol{q}^F]) - \mathrm{E}[\boldsymbol{q}^F]E[\boldsymbol{q}^F]^\dagger\right) + \left(1 - \frac{1}{\lambda N^T}\right) \mathrm{var}[\boldsymbol{q}^F] \\
&= \frac{1}{\lambda N^T} M\left(\boldsymbol{\mu}^B\right) + \left(1 - \frac{1}{\lambda N^T}\right)\left(\gamma_0 M(\boldsymbol{\mu}^B) + \delta_0 \Sigma^B\right) \\
&= \left(\frac{1}{\lambda N^T} + \left(1 - \frac{1}{\lambda N^T}\right)\gamma_0\right) M\left(\boldsymbol{\mu}^B\right) + \left(1 - \frac{1}{\lambda N^T}\right)\delta_0 \Sigma^B \\
&\equiv \gamma_{1,1} M\left(\boldsymbol{\mu}^B\right) + \delta_1 \Sigma^B
\end{aligned} \tag{D.10}$$

where in the last step we defined $\gamma_{1,1} = \frac{1}{\lambda N^T} + \left(1 - \frac{1}{\lambda N^T}\right)\gamma_0$ and $\delta_1 = \left(1 - \frac{1}{\lambda N^T}\right)\delta_0$ (to become clear later).

Continuing with the marginalisation over $\boldsymbol{q}^{\boldsymbol{G_1}}$:

$$\mathrm{E}[\boldsymbol{q}^{G_2}] = \mathrm{E}[\mathrm{E}[\boldsymbol{q}^{G_2}|\boldsymbol{q}^{G_1}]] = \mathrm{E}[S^G(\boldsymbol{q}^{G_1})] \approx S^G(\mathrm{E}[\boldsymbol{q}^{G_1}]) = S^G(\boldsymbol{\mu}^B) \tag{D.11}$$

where the approximation in the penultimate step was due to the first-order second-moment method.

Next, under the law of total variance,

$$\begin{aligned}
\mathrm{var}(\boldsymbol{q}^{G_2}) &= \mathrm{E}[\mathrm{var}[\boldsymbol{q}^{G_2}|\boldsymbol{q}^{G_1}]] + \mathrm{var}[\mathrm{E}[\boldsymbol{q}^{G_2}|\boldsymbol{q}^{G_1}]] \\
&= \mathrm{E}\left[\frac{1}{\lambda^2 N^T}\left(\mathrm{Diag}(S^G(\boldsymbol{q}^{G_1})) - S^G(\boldsymbol{q}^{G_1})(S^G(\boldsymbol{q}^{G_1}))^\dagger\right)\right] + \mathrm{var}[S^G(\boldsymbol{q}^{G_1})] \\
&= \frac{1}{\lambda^2 N^T}\left(\mathrm{Diag}(\mathrm{E}[S^G(\boldsymbol{q}^{G_1})]) - \mathrm{E}[S^G(\boldsymbol{q}^{G_1})]E[S^G(\boldsymbol{q}^{G_1})]^\dagger\right) \\
&\qquad + \left(1 - \frac{1}{\lambda^2 N^T}\right)\mathrm{var}[S^G(\boldsymbol{q}^{G_1})] \\
&= \frac{1}{\lambda^2 N^T} M\left(\mathrm{E}[S^G(\boldsymbol{q}^{G_1})]\right) + \left(1 - \frac{1}{\lambda^2 N^T}\right)\mathrm{var}[S^G(\boldsymbol{q}^{G_1})] \\
&\approx \frac{1}{\lambda^2 N^T} M\left(S^G(\mathrm{E}[\boldsymbol{q}^{G_1}])\right) \\
&\qquad + \left(1 - \frac{1}{\lambda^2 N^T}\right)\left(DS^G\big|_{\mathrm{E}[\boldsymbol{q}^{G_1}]}\right)\mathrm{var}\left[\boldsymbol{q}^{G_1}\right]\left(DS^G\big|_{\mathrm{E}[\boldsymbol{q}^{G_1}]}\right)^\dagger \\
&= \frac{1}{\lambda^2 N^T} M\left(S^G(\boldsymbol{\mu}^B)\right) + \left(1 - \frac{1}{\lambda^2 N^T}\right)\left(DS^G\big|_{\boldsymbol{\mu}^B}\right)\left(\gamma_{1,1} M(\boldsymbol{\mu}^B)\right. \\
&\qquad \left. + \delta_1 \Sigma^B\right)\left(DS^G\big|_{\boldsymbol{\mu}^B}\right)^\dagger \\
&= \frac{1}{\lambda^2 N^T} M\left(S^G(\boldsymbol{\mu}^B)\right) + \left(1 - \frac{1}{\lambda^2 N^T}\right)\gamma_{1,1}\left(DS^G\big|_{\boldsymbol{\mu}^B}\right) M(\boldsymbol{\mu}^B)\left(DS^G\big|_{\boldsymbol{\mu}^B}\right)^\dagger \\
&\qquad + \left(1 - \frac{1}{\lambda^2 N^T}\right)\delta_1\left(DS^G\big|_{\boldsymbol{\mu}^B}\right)\Sigma^B\left(DS^G\big|_{\boldsymbol{\mu}^B}\right)^\dagger \\
&= \gamma_{2,2} M\left(S^G(\boldsymbol{\mu}^B)\right) + \gamma_{2,1}\left(DS^G\big|_{\boldsymbol{\mu}^B}\right) M(\boldsymbol{\mu}^B)\left(DS^G\big|_{\boldsymbol{\mu}^B}\right)^\dagger \\
&\qquad + \delta_2\left(DS^G\big|_{\boldsymbol{\mu}^B}\right)\Sigma^B\left(DS^G\big|_{\boldsymbol{\mu}^B}\right)^\dagger
\end{aligned} \tag{D.12}$$

where we defined $\gamma_{2,1} = \left(1 - \frac{1}{\lambda^2 N^T}\right)\gamma_{1,1}$, $\gamma_{2,2} = \frac{1}{\lambda^2 N^T}$ and $\delta_2 = \left(1 - \frac{1}{\lambda^2 N^T}\right)\delta_1$.

We notice that we can define the mean and variance of an arbitrary time step n as

$$\mathrm{E}[\boldsymbol{q}^{G_n}] = (S^G)^{n-1}(\boldsymbol{\mu}^B) \tag{D.13}$$

and

$$\begin{aligned}\mathrm{var}(\boldsymbol{q}^{G_n}) = \delta_n \left(DS^G\big|_{\boldsymbol{\mu}^B}\right)^{n-1} \Sigma^B \left(\left(DS^G\big|_{\boldsymbol{\mu}^B}\right)^{\dagger}\right)^{n-1} \\ + \sum_{j=1}^{n} \gamma_{n,j} \left(DS^G\big|_{\boldsymbol{\mu}^B}\right)^{n-j} M\left((S^G)^{j-1}(\boldsymbol{\mu}^B)\right) \left(\left(DS^G\big|_{\boldsymbol{\mu}^B}\right)^{\dagger}\right)^{n-j}\end{aligned} \tag{D.14}$$

where $(S^G)^k(\boldsymbol{\mu}^B)$ denotes k applications of S^G, e.g.

$$(S^G)^3(\boldsymbol{\mu}^B) = S^G(S^G(S^G(\boldsymbol{\mu}^B))) \neq (S^G(\boldsymbol{\mu}^B))^3 \tag{D.15}$$

and $\left(DS^G\big|_{\boldsymbol{\mu}^B}\right)^k$ describes the product of k Jacobian matrices,

$$\left(DS^G\big|_{\boldsymbol{\mu}^B}\right)^k = \prod_{j=1}^{k} \left(DS^G\big|_{(S^G)^{k-j}(\boldsymbol{\mu}^B)}\right) \tag{D.16}$$

take for instance

$$\left(DS^G\big|_{\boldsymbol{\mu}^B}\right)^3 = \left(DS^G\big|_{(S^G)^2(\boldsymbol{\mu}^B)}\right)\left(DS^G\big|_{(S^G)(\boldsymbol{\mu}^B)}\right)\left(DS^G\big|_{\boldsymbol{\mu}^B}\right) \tag{D.17}$$

We define $(S^G)^0(\boldsymbol{\mu}^B) = \boldsymbol{\mu}^B$ and $\left(DS^G\big|_{\boldsymbol{\mu}^B}\right)^0 = \mathbb{1}$ where $\mathbb{1}$ is the identity matrix.

The coefficients $\gamma_{n,j}$ and δ_n obey the recurrence relations:

$$\gamma_{n,j} = \begin{cases} \left(1 - \frac{1}{\lambda^n N^t}\right)\gamma_{n-1,j}, & \text{if } j < n \\ \frac{1}{\lambda^n N^T}, & \text{if } j = n \end{cases}$$

$$\delta_n = \left(1 - \frac{1}{\lambda^n N^t}\right)\delta_{n-1} \tag{D.18}$$

for $n > 1$ with $\gamma_{1,1} = \frac{1}{\lambda N^T} + \left(1 - \frac{1}{\lambda N^T}\right)\frac{1}{N^T}$ and $\delta_1 = \left(1 - \frac{1}{\lambda N^T}\right)\left(1 - \frac{1}{N^T}\right)$.

The marginalisation over $\boldsymbol{q}^A$ may be performed in a manner similar to that employed in deriving Equations 4.6 and 4.7.

D.3 Selection for Transmission and Within-Host Adaptation

In this section we consider selection both for increased transmissibility and for within-host adaptation. First we outline the conditional distributions required for computing the compound solutions. Here we have a founder population characterised by selection for transmission:

$$\boldsymbol{q}^F \sim \mathcal{N}\left(S^T(\boldsymbol{q}^B), \tfrac{1}{N^T} M(S^T(\boldsymbol{q}^B))\right) \tag{D.19}$$

where $\mathcal{N}$ denotes the multivariate normal distribution.

In terms of viral growth, we first have a round of expansion during which no selection applies:

$$\boldsymbol{q}^{G_1} \sim \mathcal{N}\left(\boldsymbol{q}^F, \tfrac{1}{N^{G_1}} M(\boldsymbol{q}^F)\right) = \mathcal{N}\left(\boldsymbol{q}^F, \tfrac{1}{\lambda N^T} M(\boldsymbol{q}^F)\right) \tag{D.20}$$

where we defined $N^{G_1} = \lambda N^T$.

Next we have a growth round during which selection acts (prior to growth):

$$\boldsymbol{q}^{G_2} \sim \mathcal{N}\left(S^G(\boldsymbol{q}^{G_1}), \tfrac{1}{N^{G_2}} M(S^G(\boldsymbol{q}^{G_1}))\right) = \mathcal{N}\left(S^G(\boldsymbol{q}^{G_1}), \tfrac{1}{\lambda^2 N^T} M(S^G(\boldsymbol{q}^{G_1}))\right) \tag{D.21}$$

where S^G is in 12-hour units and $N^{G_2} = \lambda^2 N^T$.

In the same manner, every subsequent growth round will incorporate both growth and selection with selection acting first. In general we have that

$$\begin{aligned} \boldsymbol{q}^{G_n} &\sim \mathcal{N}\left(S^G(\boldsymbol{q}^{G_{n-1}}), \tfrac{1}{N^{G_n}} M(S^G(\boldsymbol{q}^{G_{n-1}}))\right) \\ &= \mathcal{N}\left(S^G(\boldsymbol{q}^{G_{n-1}}), \tfrac{1}{\lambda^n N^T} M(S^G(\boldsymbol{q}^{G_{n-1}}))\right) \end{aligned} \tag{D.22}$$

with $N^{G_n} = \lambda^n N^T$ for $n > 1$.

Assuming N steps in the growth process (e.g. if one step = 12 hours, then $N = 2$ steps would correspond to a 24 hour difference between donor and recipient sampling times), we have

$$\boldsymbol{q}^A \sim \mathcal{N}\left(S^G(\boldsymbol{q}^{G_{N-1}}), \tfrac{1}{N^{G_N}} M(S^G(\boldsymbol{q}^{G_{N-1}}))\right) \tag{D.23}$$

Finally, as always, the post-transmission sampling step is defined by

$$\boldsymbol{x}_i^{A,\mathrm{P}} \sim \mathcal{N}\left(N_i^A T_i S^G(\boldsymbol{q}^A), \alpha_i N_i^A M(T_i S^G(\boldsymbol{q}^A)\right) \tag{D.24}$$

where $\alpha_i = \frac{N_i^A + C}{1+C}$. Note that selection acts prior to sampling here. This ensures that there are equally many drift and selection steps during viral growth.

Turning now to the evaluation of compound distributions, the marginalisation over $\boldsymbol{q}^B$ leads to

$$\mathrm{E}[\boldsymbol{q}^F] = \mathrm{E}[\mathrm{E}[\boldsymbol{q}^F|\boldsymbol{q}^B]] = \mathrm{E}[S^T(\boldsymbol{q}^B)] \approx S^T(\mathrm{E}[\boldsymbol{q}^B]) = S^T(\boldsymbol{\mu}^B) \tag{D.25}$$

where in the penultimate step we used the first-order second-moment approximation to a vector function acting on a random variable. The law of total variance yields

$$\begin{aligned}
\mathrm{var}(\boldsymbol{q}^F) &= \mathrm{E}[\mathrm{var}[\boldsymbol{q}^F|\boldsymbol{q}^B]] + \mathrm{var}[\mathrm{E}[\boldsymbol{q}^F|\boldsymbol{q}^B]] \\
&= \mathrm{E}\left[\frac{1}{N^T} M(S^T(\boldsymbol{q}^B))\right] + \mathrm{var}\left[S^T(\boldsymbol{q}^B)\right] \\
&= \frac{1}{N^T} M(\mathrm{E}[S^T(\boldsymbol{q}^B)]) + \left(1 - \frac{1}{N^T}\right) \mathrm{var}[S^T(\boldsymbol{q}^B)] \\
&\approx \frac{1}{N^T} M(S^T(\mathrm{E}[\boldsymbol{q}^B])) + \left(1 - \frac{1}{N^T}\right) \left(DS^T\big|_{\mathrm{E}[\boldsymbol{q}^B]}\right) \mathrm{var}[\boldsymbol{q}^B] \left(DS^T\big|_{\mathrm{E}[\boldsymbol{q}^B]}\right)^\dagger \\
&= \frac{1}{N^T} M(S^T(\boldsymbol{\mu}^B)) + \left(1 - \frac{1}{N^T}\right) \left(DS^T\big|_{\boldsymbol{\mu}^B}\right) \Sigma^B \left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger \\
&= \gamma_0 M(S^T(\boldsymbol{\mu}^B)) + \delta_0 \left(DS^T\big|_{\boldsymbol{\mu}^B}\right) \Sigma^B \left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger
\end{aligned} \tag{D.26}$$

where in the last step we defined $\gamma_0 = \frac{1}{N^T}$ and $\delta_0 = \left(1 - \frac{1}{N^T}\right)$. We also note that $(DS)_i^j = \frac{\partial S_i}{\partial q_j}$ is the Jacobian matrix arising from the first-order second-moment approximation.

Next, for the $\boldsymbol{q}^F$ integral, the law of total expectation yields

$$\mathrm{E}[\boldsymbol{q}^{G_1}] = \mathrm{E}[\mathrm{E}[\boldsymbol{q}^{G_1}|\boldsymbol{q}^F]] = \mathrm{E}[\boldsymbol{q}^F] = S^T(\boldsymbol{\mu}^B) \tag{D.27}$$

whilst the law of total variance gives

$$\begin{aligned}
\mathrm{var}(\boldsymbol{q}^{G_1}) &= \mathrm{E}[\mathrm{var}[\boldsymbol{q}^{G_1}|\boldsymbol{q}^F]] + \mathrm{var}[\mathrm{E}[\boldsymbol{q}^{G_1}|\boldsymbol{q}^F]] \\
&= \mathrm{E}\left[\frac{1}{\lambda N^T}\left(\mathrm{Diag}(\boldsymbol{q}^F) - \boldsymbol{q}^F(\boldsymbol{q}^F)^\dagger\right)\right] + \mathrm{var}[\boldsymbol{q}^F] \\
&= \frac{1}{\lambda N^T}\left(\mathrm{Diag}(\mathrm{E}[\boldsymbol{q}^F]) - \mathrm{E}[\boldsymbol{q}^F]E[\boldsymbol{q}^F]^\dagger\right) + \left(1 - \frac{1}{\lambda N^T}\right)\mathrm{var}[\boldsymbol{q}^F] \\
&= \frac{1}{\lambda N^T} M\left(S^T(\boldsymbol{\mu}^B)\right) + \left(1 - \frac{1}{\lambda N^T}\right)\left(\gamma_0 M(S^T(\boldsymbol{\mu}^B))\right. \\
&\qquad \left. + \delta_0 \left(DS^T\big|_{\boldsymbol{\mu}^B}\right)\Sigma^B\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger\right) \\
&= \left(\frac{1}{\lambda N^T} + \left(1 - \frac{1}{\lambda N^T}\right)\gamma_0\right) M\left(S^T(\boldsymbol{\mu}^B)\right) \\
&\qquad + \left(1 - \frac{1}{\lambda N^T}\right)\delta_0\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)\Sigma^B\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger \\
&\equiv \gamma_{1,1} M\left(S^T(\boldsymbol{\mu}^B)\right) + \delta_1\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)\Sigma^B\left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger
\end{aligned} \tag{D.28}$$

where in the last step we defined $\gamma_{1,1} = \frac{1}{\lambda N^T} + \left(1 - \frac{1}{\lambda N^T}\right)\gamma_0$ and $\delta_1 = \left(1 - \frac{1}{\lambda N^T}\right)\delta_0$ (to become clear later).

Continuing with the marginalisation over $\boldsymbol{q}^{\boldsymbol{G_1}}$:

$$\mathrm{E}[\boldsymbol{q}^{G_2}] = \mathrm{E}[\mathrm{E}[\boldsymbol{q}^{G_2}|\boldsymbol{q}^{G_1}]] = \mathrm{E}[S^G(\boldsymbol{q}^{G_1})] \approx S^G(\mathrm{E}[\boldsymbol{q}^{G_1}]) = S^G(S^T(\boldsymbol{\mu}^B)) \tag{D.29}$$

where the approximation in the penultimate step was due to the first-order second-moment method.

Next, under the law of total variance,

$$\begin{aligned}
\mathrm{var}(\boldsymbol{q}^{G_2}) &= \mathrm{E}[\mathrm{var}[\boldsymbol{q}^{G_2}|\boldsymbol{q}^{G_1}]] + \mathrm{var}[\mathrm{E}[\boldsymbol{q}^{G_2}|\boldsymbol{q}^{G_1}]] \\
&= \mathrm{E}\left[\frac{1}{\lambda^2 N^T}\left(\mathrm{Diag}(S^G(\boldsymbol{q}^{G_1})) - S^G(\boldsymbol{q}^{G_1})(S^G(\boldsymbol{q}^{G_1}))^\dagger\right)\right] + \mathrm{var}[S^G(\boldsymbol{q}^{G_1})] \\
&= \frac{1}{\lambda^2 N^T}\left(\mathrm{Diag}(\mathrm{E}[S^G(\boldsymbol{q}^{G_1})]) - \mathrm{E}[S^G(\boldsymbol{q}^{G_1})]E[S^G(\boldsymbol{q}^{G_1})]^\dagger\right) \\
&\qquad + \left(1 - \frac{1}{\lambda^2 N^T}\right)\mathrm{var}[S^G(\boldsymbol{q}^{G_1})] \\
&= \frac{1}{\lambda^2 N^T} M\left(\mathrm{E}[S^G(\boldsymbol{q}^{G_1})]\right) + \left(1 - \frac{1}{\lambda^2 N^T}\right)\mathrm{var}[S^G(\boldsymbol{q}^{G_1})]
\end{aligned} \tag{D.30}$$

$$
\begin{aligned}
&\approx \frac{1}{\lambda^2 N^T} M\left(S^G(\mathrm{E}[\boldsymbol{q}^{G_1}])\right) \\
&\qquad + \left(1 - \frac{1}{\lambda^2 N^T}\right) \left(DS^G\big|_{\mathrm{E}[\boldsymbol{q}^{G_1}]}\right) \mathrm{var}\left[\boldsymbol{q}^{G_1}\right] \left(DS^G\big|_{\mathrm{E}[\boldsymbol{q}^{G_1}]}\right)^\dagger \\
&= \frac{1}{\lambda^2 N^T} M\left(S^G(S^T(\boldsymbol{\mu}^B))\right) \\
&\qquad + \left(1 - \frac{1}{\lambda^2 N^T}\right) \left(DS^G\big|_{S^T(\boldsymbol{\mu}^B)}\right) \Big(\gamma_{1,1} M\left(S^T(\boldsymbol{\mu}^B)\right) \\
&\qquad\qquad + \delta_1 \left(DS^T\big|_{\boldsymbol{\mu}^B}\right) \Sigma^B \left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger\Big) \left(DS^G\big|_{S^T(\boldsymbol{\mu}^B)}\right)^\dagger \\
&= \frac{1}{\lambda^2 N^T} M\left(S^G(S^T(\boldsymbol{\mu}^B))\right) \\
&\qquad + \left(1 - \frac{1}{\lambda^2 N^T}\right) \gamma_{1,1} \left(DS^G\big|_{S^T(\boldsymbol{\mu}^B)}\right) M(S^T(\boldsymbol{\mu}^B) \left(DS^G\big|_{S^T(\boldsymbol{\mu}^B)}\right)^\dagger \\
&\qquad + \left(1 - \frac{1}{\lambda^2 N^T}\right) \delta_1 \left(DS^G\big|_{S^T(\boldsymbol{\mu}^B)}\right) \left(DS^T\big|_{\boldsymbol{\mu}^B}\right) \Sigma^B \\
&\qquad\qquad \times \left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger \left(DS^G\big|_{S^T(\boldsymbol{\mu}^B)}\right)^\dagger \\
&= \gamma_{2,2} M\left(S^G(S^T(\boldsymbol{\mu}^B))\right) + \gamma_{2,1} \left(DS^G\big|_{S^T(\boldsymbol{\mu}^B)}\right) M(S^T(\boldsymbol{\mu}^B)) \left(DS^G\big|_{S^T(\boldsymbol{\mu}^B)}\right)^\dagger \\
&\qquad + \delta_2 \left(DS^G\big|_{S^T(\boldsymbol{\mu}^B)}\right) \left(DS^T\big|_{\boldsymbol{\mu}^B}\right) \Sigma^B \left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger \left(DS^G\big|_{S^T(\boldsymbol{\mu}^B)}\right)^\dagger
\end{aligned}
$$

where we defined $\gamma_{2,1} = \left(1 - \frac{1}{\lambda^2 N^T}\right) \gamma_{1,1}$, $\gamma_{2,2} = \frac{1}{\lambda^2 N^T}$ and $\delta_2 = \left(1 - \frac{1}{\lambda^2 N^T}\right) \delta_1$.

We notice that we can define the mean and variance of an arbitrary time step n as

$$
\mathrm{E}[\boldsymbol{q}^{G_n}] = (S^G)^{n-1}(S^T(\boldsymbol{\mu}^B)) \tag{D.31}
$$

and

$$
\begin{aligned}
\mathrm{var}(\boldsymbol{q}^{G_n}) = \delta_n &\left(DS^G\big|_{S^T(\boldsymbol{\mu}^B)}\right)^{n-1} \left(DS^T\big|_{\boldsymbol{\mu}^B}\right) \Sigma^B \left(DS^T\big|_{\boldsymbol{\mu}^B}\right)^\dagger \left(\left(DS^G\big|_{S^T(\boldsymbol{\mu}^B)}\right)^\dagger\right)^{n-1} \\
&+ \sum_{j=1}^{n} \gamma_{n,j} \left(DS^G\big|_{S^T(\boldsymbol{\mu}^B)}\right)^{n-j} M\left((S^G)^{j-1}(S^T(\boldsymbol{\mu}^B))\right) \left(\left(DS^G\big|_{S^T(\boldsymbol{\mu}^B)}\right)^\dagger\right)^{n-j}
\end{aligned} \tag{D.32}
$$

where $(S^G)^k(S^T(\boldsymbol{\mu}^B))$ denotes k applications of S^G, e.g.

$$(S^G)^3(S^T(\boldsymbol{\mu}^B)) = S^G(S^G(S^G(S^T(\boldsymbol{\mu}^B)))) \neq (S^G(S^T(\boldsymbol{\mu}^B))^3 \tag{D.33}$$

and $\left(DS^G\big|_{S^T(\boldsymbol{\mu}^B)}\right)^k$ describes the product of k Jacobian matrices,

$$\left(DS^G\big|_{S^T(\boldsymbol{\mu}^B)}\right)^k = \prod_{j=1}^{k}\left(DS^G\big|_{(S^G)^{k-j}(S^T(\boldsymbol{\mu}^B))}\right) \tag{D.34}$$

take for instance

$$\left(DS^G\big|_{S^T(\boldsymbol{\mu}^B)}\right)^3 = \left(DS^G\big|_{(S^G)^2(S^T(\boldsymbol{\mu}^B))}\right)\left(DS^G\big|_{(S^G)(S^T(\boldsymbol{\mu}^B))}\right)\left(DS^G\big|_{S^T(\boldsymbol{\mu}^B)}\right) \tag{D.35}$$

We define $(S^G)^0(S^T(\boldsymbol{\mu}^B)) = S^T(\boldsymbol{\mu}^B)$ and $\left(DS^G\big|_{S^T(\boldsymbol{\mu}^B)}\right)^0 = \mathbb{1}$ where $\mathbb{1}$ is the identity matrix.

The coefficients $\gamma_{n,j}$ and δ_n obey the recurrence relations:

$$\gamma_{n,j} = \begin{cases} \left(1 - \frac{1}{\lambda^n N^t}\right)\gamma_{n-1,j}, & \text{if } j < n \\ \frac{1}{\lambda^n N^T}, & \text{if } j = n \end{cases}$$

$$\delta_n = \left(1 - \frac{1}{\lambda^n N^t}\right)\delta_{n-1} \tag{D.36}$$

for $n > 1$ with $\gamma_{1,1} = \frac{1}{\lambda N^T} + \left(1 - \frac{1}{\lambda N^T}\right)\frac{1}{N^T}$ and $\delta_1 = \left(1 - \frac{1}{\lambda N^T}\right)\left(1 - \frac{1}{N^T}\right)$.

The marginalisation over $\boldsymbol{q}^A$ may be performed in a manner similar to that employed in deriving Equations 4.6 and 4.7.

Appendix E

Proof that $T\mathrm{Diag}(\boldsymbol{q})T^{\dagger} = \mathrm{Diag}(T\boldsymbol{q})$

E.1 Introduction

In our derivations of compound distributions we use that the identity $T\mathrm{Diag}(\boldsymbol{q})T^{\dagger} = \mathrm{Diag}(T\boldsymbol{q})$ is true for a $J \times K$ matrix T and a K dimensional vector $\boldsymbol{q}$ if T consists of zeroes and ones and if every column of T contains a single non-zero element, i.e. if a full haplotype can only contribute to a single partial haplotype in the partial haplotype set. Here we have suppressed the subscripts denoting partial haplotype sets to avoid confusion in the subsequent derivation.

E.2 Proof

Considering first the right hand side of the identity, we see that

$$\mathrm{Diag}(T\boldsymbol{q})_{i,j} = \begin{cases} T_{i,k}q_k, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$$

where we have used implicit summation over the k index.

Considering the left hand side of the identity, we may examine the cases of $i = j$ and $i \neq j$ separately. For $i = j$ we can write the left hand side as

$$\left(T\mathrm{Diag}(\boldsymbol{q})T^{\dagger}\right)_{i,i} = T_{i,k}\mathrm{Diag}(\boldsymbol{q})_{k,l}T^{\dagger}_{l,i} = T_{i,k}\delta_{k,l}q_l T^{\dagger}_{l,i} \tag{E.1}$$

where there is no summation over the i indices. Additionally, we have represented $\mathrm{Diag}(\boldsymbol{q})_{k,l}$ as $\delta_{k,l}q_l$ where $\delta_{k,l}$ is the Kronecker delta/the identity matrix. Here there is no summation over l, even though, slightly confusingly, the sum-

mation over l is implied in $\text{Diag}(\boldsymbol{q})_{k,l}T^\dagger_{l,i}$. In other words, q_l may be thought of as an index-valued scaling factor to the summed over matrix $\delta_{k,l}$.

Using $T^\dagger_{l,i} = T_{i,l}$ and the replacement properties of the Kronecker delta yields

$$\left(T\text{Diag}(\boldsymbol{q})T^\dagger\right)_{i,i} = T_{i,k}\delta_{k,l}q_l T_{i,l} = T_{i,k}q_k T_{i,k} = q_k \left(T_{i,k}\right)^2 \tag{E.2}$$

As all entries of T are zeroes and ones, we must have that $\left(T_{i,k}\right)^2 = T_{i,k}$. Thus,

$$\left(T\text{Diag}(\boldsymbol{q})T^\dagger\right)_{i,i} = T_{i,k}q_k \tag{E.3}$$

as required.

Considering the case of $i \neq j$:

$$\left(T\text{Diag}(\boldsymbol{q})T^\dagger\right)_{i,j} = T_{i,k}\text{Diag}(\boldsymbol{q})_{k,l}T^\dagger_{l,j} = T_{i,k}\delta_{k,l}q_l T^\dagger_{l,j} = T_{i,k}\delta_{k,l}q_l T_{j,l} = T_{i,k}q_k T_{j,k} \tag{E.4}$$

Now we use the property that each column of T must have exactly one non-zero entry. This implies that $T_{i,k}T_{j,k} = 0$ if $i \neq j$. Thus, for $i \neq j$,

$$\left(T\text{Diag}(\boldsymbol{q})T^\dagger\right)_{i,j} = 0 \tag{E.5}$$

as required. This proves the identity.

Appendix F

First-Order Second-Moment Method for Vector Functions

F.1 Introduction

Consider a random vector $\boldsymbol{X}$ with probability density function $f_X(\boldsymbol{x})$ and realisation $\boldsymbol{x} \in \mathbb{R}^k$, then, given a vector-valued function $\boldsymbol{g}(\boldsymbol{x}) : \mathbb{R}^k \to \mathbb{R}^k$, it is of interest to find approximations to the mean, $\mathrm{E}[\boldsymbol{g}(\boldsymbol{x})]$, and variance, $\mathrm{var}[\boldsymbol{g}(\boldsymbol{x})]$, of the function.

F.2 Derivation

The function of interest may be approximated by a Taylor expansion of the form

$$\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{\mu}) + D\boldsymbol{g}(\boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu}) + O(\boldsymbol{x}^2) \tag{F.1}$$

where $\boldsymbol{\mu}$ is the mean of $\boldsymbol{X}$ and $D\boldsymbol{g}(\boldsymbol{x})_i^j = \frac{\partial g_i}{\partial x_j}$ is the Jacobian matrix.

The mean of $\boldsymbol{g}(\boldsymbol{x})$ is given by

$$\boldsymbol{\mu}_g = \mathrm{E}[\boldsymbol{g}(\boldsymbol{x})] = \int_{-\infty}^{\infty} \boldsymbol{g}(\boldsymbol{x}) f_X(\boldsymbol{x}) d\boldsymbol{x} \tag{F.2}$$

Approximating the mean by the Taylor expansion to first order yields

$$\begin{aligned}
\boldsymbol{\mu}_g &\approx \int_{-\infty}^{\infty} \left[\boldsymbol{g}(\boldsymbol{\mu}) + D\boldsymbol{g}(\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})\right] f_X(\boldsymbol{x}) d\boldsymbol{x} \\
&= \int_{-\infty}^{\infty} \boldsymbol{g}(\boldsymbol{\mu}) f_X(\boldsymbol{x}) d\boldsymbol{x} + \int_{-\infty}^{\infty} D\boldsymbol{g}(\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu}) f_X(\boldsymbol{x}) d\boldsymbol{x} \\
&= \boldsymbol{g}(\boldsymbol{\mu}) \underbrace{\int_{-\infty}^{\infty} f_X(\boldsymbol{x}) d\boldsymbol{x}}_{1} + D\boldsymbol{g}(\boldsymbol{\mu}) \underbrace{\int_{-\infty}^{\infty} (\boldsymbol{x}-\boldsymbol{\mu}) f_X(\boldsymbol{x}) d\boldsymbol{x}}_{0} \\
&= \boldsymbol{g}(\boldsymbol{\mu})
\end{aligned} \tag{F.3}$$

This is the first order approximation to the mean.

The variance is given by

$$\operatorname{var}\left[\boldsymbol{g}(\boldsymbol{x})\right] = \mathrm{E}\left[\boldsymbol{g}(\boldsymbol{x})^2\right] - \boldsymbol{\mu}_g^2 = \int_{-\infty}^{\infty} \boldsymbol{g}(\boldsymbol{x})^2 f(\boldsymbol{x}) d\boldsymbol{x} - \boldsymbol{\mu}_g^2 \tag{F.4}$$

Inserting the Taylor expansion (to first order) gives

$$\begin{aligned}
\operatorname{var}\left[\boldsymbol{g}(\boldsymbol{x})\right] &\approx \int_{-\infty}^{\infty} \left(\boldsymbol{g}(\boldsymbol{\mu}) + D\boldsymbol{g}(\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})\right)^2 f(\boldsymbol{x}) d\boldsymbol{x} - \boldsymbol{\mu}_g^2 \\
&= \int_{-\infty}^{\infty} \left(\boldsymbol{g}(\boldsymbol{\mu}) + D\boldsymbol{g}(\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})\right) \left(\boldsymbol{g}(\boldsymbol{\mu})\right. \\
&\qquad \left. + D\boldsymbol{g}(\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})\right)^{\dagger} f(\boldsymbol{x}) d\boldsymbol{x} - \boldsymbol{\mu}_g^2 \\
&= \int_{-\infty}^{\infty} \left(\boldsymbol{g}(\boldsymbol{\mu}) + D\boldsymbol{g}(\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})\right) \left(\boldsymbol{g}(\boldsymbol{\mu})^{\dagger}\right. \\
&\qquad \left. + (\boldsymbol{x}-\boldsymbol{\mu})^{\dagger} D\boldsymbol{g}(\boldsymbol{\mu})^{\dagger}\right) f(\boldsymbol{x}) d\boldsymbol{x} - \boldsymbol{\mu}_g^2 \\
&= \int_{-\infty}^{\infty} \left(\boldsymbol{g}(\boldsymbol{\mu})^2 + D\boldsymbol{g}(\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})\boldsymbol{g}(\boldsymbol{\mu})^{\dagger} + \boldsymbol{g}(\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\dagger} D\boldsymbol{g}(\boldsymbol{\mu})^{\dagger}\right. \\
&\qquad \left. + D\boldsymbol{g}(\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^2 D\boldsymbol{g}(\boldsymbol{\mu})^{\dagger}\right) f(\boldsymbol{x}) d\boldsymbol{x} - \boldsymbol{\mu}_g^2 \\
&= \boldsymbol{g}(\boldsymbol{\mu})^2 \underbrace{\int_{-\infty}^{\infty} f_X(\boldsymbol{x}) d\boldsymbol{x}}_{1} + D\boldsymbol{g}(\boldsymbol{\mu}) \underbrace{\int_{-\infty}^{\infty} (\boldsymbol{x}-\boldsymbol{\mu}) f_X(\boldsymbol{x}) d\boldsymbol{x}}_{0} \boldsymbol{g}(\boldsymbol{\mu})^{\dagger} \\
&\qquad + \boldsymbol{g}(\boldsymbol{\mu}) \underbrace{\int_{-\infty}^{\infty} (\boldsymbol{x}-\boldsymbol{\mu}) f_X(\boldsymbol{x}) d\boldsymbol{x}}_{0} D\boldsymbol{g}(\boldsymbol{\mu})^{\dagger} \\
&\qquad + D\boldsymbol{g}(\boldsymbol{\mu}) \underbrace{\int_{-\infty}^{\infty} (\boldsymbol{x}-\boldsymbol{\mu})^2 f_X(\boldsymbol{x}) d\boldsymbol{x}}_{\operatorname{cov}[\boldsymbol{x},\boldsymbol{x}] \equiv \Sigma} D\boldsymbol{g}(\boldsymbol{\mu})^{\dagger} - \boldsymbol{\mu}_g^2
\end{aligned} \tag{F.5}$$

$$
\begin{aligned}
&= \boldsymbol{g}(\boldsymbol{\mu})^2 + D\boldsymbol{g}(\boldsymbol{\mu})\Sigma D\boldsymbol{g}(\boldsymbol{\mu})^\dagger - \boldsymbol{\mu}_g^2 \\
&\approx \boldsymbol{\mu}_g^2 + D\boldsymbol{g}(\boldsymbol{\mu})\Sigma D\boldsymbol{g}(\boldsymbol{\mu})^\dagger - \boldsymbol{\mu}_g^2 \\
&= D\boldsymbol{g}(\boldsymbol{\mu})\Sigma D\boldsymbol{g}(\boldsymbol{\mu})^\dagger
\end{aligned}
$$

This is the first order approximation to the variance.