Leveraging genetic time series data to improve detection of natural selection

Luvenia Nicole Hellams
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd
Part of the Statistics and Probability Commons

Recommended Citation
https://lib.dr.iastate.edu/etd/16372

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Leveraging genetic time series data to improve detection of natural selection

by

Luvenia Nicole Hellams

A dissertation submitted to the graduate faculty in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Karin Dorman, Major Professor
Susan Carpenter
Daniel Nettleton
Jarad Niemi
Daniel Nordman

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University
Ames, Iowa
2018

Copyright © Luvenia Nicole Hellams, 2018. All rights reserved.
In loving memory of
Aunt Velma, Uncle Herman, Aunt Naomi and Aunt Veronica

The reality of higher education is that racial and gender biases in academia are real and black women lie delicately in the intersection of the two. Nothing gives me more joy than the unwavering determination of black women scholars in spite of the fact that everything in the creation of these institutions demands and encourages our demise. In a time when anti-blackness is venomously inescapable, even within our institutions of higher education, may black scholarship be infinite.
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>viii</td>
</tr>
<tr>
<td>CHAPTER 1. LITERATURE REVIEW</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Biological Primer</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Selection and Theoretical Population Genetics</td>
<td>3</td>
</tr>
<tr>
<td>1.2.1 Wright-Fisher Model</td>
<td>3</td>
</tr>
<tr>
<td>1.2.2 Wright-Fisher Diffusion</td>
<td>4</td>
</tr>
<tr>
<td>1.2.3 Hidden Markov Model for Observational Data</td>
<td>5</td>
</tr>
<tr>
<td>1.2.4 Gaussian Approximation to the WF Diffusion</td>
<td>5</td>
</tr>
<tr>
<td>1.3 Estimating/Testing Selection</td>
<td>6</td>
</tr>
<tr>
<td>1.3.1 Bollback et al. (2008)</td>
<td>6</td>
</tr>
<tr>
<td>1.3.2 Frequency Increment Test</td>
<td>7</td>
</tr>
<tr>
<td>1.3.3 Frequency Increment Test with Reference Loci</td>
<td>8</td>
</tr>
<tr>
<td>1.4 Outline of dissertation</td>
<td>8</td>
</tr>
<tr>
<td>CHAPTER 2. THE CHALLENGE OF DETECTING SELECTION IN PRRSV</td>
<td>10</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>10</td>
</tr>
<tr>
<td>2.2 Methods</td>
<td>11</td>
</tr>
<tr>
<td>2.2.1 Samples, sequences, alignments and SNVs</td>
<td>11</td>
</tr>
<tr>
<td>2.2.2 AMOVA</td>
<td>14</td>
</tr>
<tr>
<td>2.2.3 Testing association of alleles and time</td>
<td>15</td>
</tr>
<tr>
<td>2.2.4 Hellams Test</td>
<td>16</td>
</tr>
</tbody>
</table>
2.3 Results

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.1 Genetic diversity</td>
<td>17</td>
</tr>
<tr>
<td>2.3.2 Genetic structure</td>
<td>19</td>
</tr>
<tr>
<td>2.3.3 Testing for association with time</td>
<td>21</td>
</tr>
<tr>
<td>2.3.4 Hellams test</td>
<td>26</td>
</tr>
<tr>
<td>2.3.5 Linkage disequilibrium</td>
<td>28</td>
</tr>
</tbody>
</table>

2.4 Discussion 

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>31</td>
</tr>
<tr>
<td>3.2 Model</td>
<td>32</td>
</tr>
<tr>
<td>3.2.1 Transformations</td>
<td>33</td>
</tr>
<tr>
<td>3.2.2 Simulation</td>
<td>35</td>
</tr>
<tr>
<td>3.2.3 Normalization with constant initial allele frequency ( x_r )</td>
<td>37</td>
</tr>
<tr>
<td>3.2.4 Normalization with variable initial allele frequency ( x_r )</td>
<td>38</td>
</tr>
<tr>
<td>3.2.5 ROC curves</td>
<td>39</td>
</tr>
<tr>
<td>3.3 Results</td>
<td>39</td>
</tr>
<tr>
<td>3.3.1 Normality</td>
<td>39</td>
</tr>
<tr>
<td>3.3.2 Modeling ( \lambda_{r1}, \mu_r, \sigma_r, \gamma_{ra} ) and ( \gamma_{rb} ) as functions of ( x_r )</td>
<td>41</td>
</tr>
<tr>
<td>3.3.3 Transforming datasets with mixed initial allele frequencies ( x_r )</td>
<td>42</td>
</tr>
<tr>
<td>3.3.4 Detecting selection</td>
<td>43</td>
</tr>
<tr>
<td>3.4 Discussion</td>
<td>43</td>
</tr>
</tbody>
</table>

CHAPTER 4. CONCLUSION

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Future Work</td>
<td>54</td>
</tr>
<tr>
<td>4.1.1 Conserving Normal Quantiles in the Tails</td>
<td>54</td>
</tr>
<tr>
<td>4.1.2 The Challenge of Linkage Disequilibrium</td>
<td>55</td>
</tr>
<tr>
<td>4.1.3 Alternative Models of the Distribution of Allele Frequency (DAF)</td>
<td>55</td>
</tr>
</tbody>
</table>
4.2 Discussion ............................................................... 58

BIBLIOGRAPHY .............................................................. 63
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.1</td>
<td>Clonal sequence data from PRRSV-infected pigs</td>
<td>13</td>
</tr>
<tr>
<td>Table 2.2</td>
<td>RNA sequence data from PRRSV-infected pigs</td>
<td>14</td>
</tr>
<tr>
<td>Table 2.3</td>
<td>Nested structure of PRRSV sequence data</td>
<td>20</td>
</tr>
<tr>
<td>Table 2.4</td>
<td>Test of association of SNVs and dpi in virus study</td>
<td>22</td>
</tr>
<tr>
<td>Table 2.5</td>
<td>Test of association of SNVs and dpi in host study</td>
<td>24</td>
</tr>
<tr>
<td>Table 2.6</td>
<td>Simulation of Hellams test</td>
<td>27</td>
</tr>
<tr>
<td>Table 2.7</td>
<td>Hellams test applied to host study</td>
<td>28</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>List of transformations</td>
<td>35</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>Simulation settings</td>
<td>37</td>
</tr>
<tr>
<td>Table 3.3</td>
<td>Average normality metrics pre- and post-transformation</td>
<td>40</td>
</tr>
<tr>
<td>Table 3.4</td>
<td>Normality metrics of transformed data using polynomial functions of $x_r$</td>
<td>42</td>
</tr>
<tr>
<td>Table 3.5</td>
<td>False positive rates for simulation B</td>
<td>44</td>
</tr>
<tr>
<td>Table 3.6</td>
<td>True positive rates for simulation B</td>
<td>45</td>
</tr>
<tr>
<td>Table 3.7</td>
<td>True positive rates for simulation B</td>
<td>46</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 2.1</td>
<td>Log virus load of PHGC trials 1–3</td>
<td>11</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>PRRSV env alignment</td>
<td>18</td>
</tr>
<tr>
<td>Figure 2.3</td>
<td>Temporal diversity</td>
<td>20</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Box-Cox transformation of scaled relative allele frequency changes</td>
<td>36</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Skewness and kurtosis of data pre- and post-transformation</td>
<td>48</td>
</tr>
<tr>
<td>Figure 3.3</td>
<td>Shapiro-Wilk test of normality</td>
<td>49</td>
</tr>
<tr>
<td>Figure 3.4</td>
<td>Polynomial fits of transformation parameters</td>
<td>50</td>
</tr>
<tr>
<td>Figure 3.5</td>
<td>Multinomial logistic polynomial regression for point masses</td>
<td>51</td>
</tr>
<tr>
<td>Figure 3.6</td>
<td>Empirical cdfs when transformations are polynomial functions of $x_r$</td>
<td>52</td>
</tr>
<tr>
<td>Figure 3.7</td>
<td>ROC curves show Box-Cox transformation superior</td>
<td>53</td>
</tr>
<tr>
<td>Figure 4.1</td>
<td>Harrell-Davis Quantile Estimates</td>
<td>61</td>
</tr>
<tr>
<td>Figure 4.2</td>
<td>Hellinger Distances of Approximations to the WF-IHMM DAF</td>
<td>62</td>
</tr>
</tbody>
</table>
ABSTRACT

My work focuses on the problem of detecting natural selection from genetic time series data. This dissertation is motivated by genomic sequence data from populations of the Porcine Reproductive and Respiratory Syndrome Virus (PRRSV) sampled temporally during the early stages of infection from the blood of multiple pigs. An important biological question in this context is to understand what forces drive genetic changes in the virus populations. From knowledge of when and how selection acts on these viruses, it is possible to discern how the host pig attacks the virus as well as how the virus responds. Ultimately, such knowledge can help to guide vaccination, breeding and treatment strategies, which could profoundly reduce the morbidity and economic loss wrought by this virus. Given counts $Y(t)$ of an allele at a locus observed in a sample from a population at discrete timepoints $0, t_1, t_2, \ldots$, my goal is to detect when there is evidence of selection acting on the allele. An increase $Y(t) > Y(0)$ may indicate selection for the allele, while a decrease could reveal selection against the allele, but inheritance across generations is a random process, and change is guaranteed even under neutral (no selection) conditions. The magnitude of pure genetic drift, the neutral random process that produces genetic change even in the absence of disruptive forces, is determined by the population size $N$, such that random fluctuations dominate in populations with small $N$, but completely disappear as $N \to \infty$. To demonstrate the conundrum, I implement the Cochran-Mantel-Haenszel (CMH) test to detect significant association between genetic alleles and time when multiple subjects and timepoints are available. Though the CMH test finds significant temporal trends in the PRRSV data, it cannot eliminate the possibility of pure genetic drift. I propose a novel, $N$-agnostic test for selection in such populations and demonstrate its properties in extensive simulation. Unfortunately, the test is particularly low-powered under some conditions, including those pervading the PRRSV dataset. Another test, the FITR test, requires estimation of $N$ but assumes normality of the temporal increments in relative allele frequency, which is also
not satisfied in the PRRSV data. I extend the FITR test to use normalizing transformations, which substantially extends the applicability of the test. I demonstrate that the transformations reduce the overall skewness and excess kurtosis of the original data, while better conserving the type-I error rate of the test. This work contributes one new and one improved test for detecting selection in genetic time series data that can aid in the fight against infectious disease as well as other selection-related applications.
CHAPTER 1. LITERATURE REVIEW

This chapter is an introduction to the problem of detecting genetic selection in time series data, particularly with virus genomes in mind. At the core of this dissertation, I am interested in the change in the relative frequency \(X(0)\) of an allele from one time to \(X(t)\) units later. An increase \(X(t) > X(0)\) can suggest selection for the allele, a decrease selection against the allele, but inheritance across generations is a random process such that change is guaranteed in any real (finite) biological population. So, when does change indicate selection, and when does it merely reflect randomness? That is the crux of the problem I address throughout this dissertation.

There are other challenges in detecting selection I do not address in this dissertation. Particularly for viruses, which are characterized by short, compact genomes, each nucleotide position (site) typically carries out multiple biological functions, most obvious in the regions of overlapping read frames. Such multi-functionality immediately eliminates some methods for detecting selection that rely on contrasting nonsynonymous and synonymous substitutions (Yang, 1998). Another challenge is the dependence among sites. Although many viruses recombine (Pérez-Losada et al., 2015), sites tend not to be independent, and it is not possible, as with longer genomes, to simply drop enough intervening sites to get effective independence. I acknowledge these challenges, but focus exclusively on distinguishing random from selection-driven changes in the relative allele frequency at a single site.

1.1 Biological Primer

The key to evolution is genetic variation, which is a byproduct of processes such as genetic drift, mutation, migration, natural selection, etc. Genetic variation within an evolving population is characterized by changes, both random and systematic, in the temporal frequencies of different gene variants (alleles) at multiple locations (loci) on a genome. Evolutionary processes do not act in
isolation, therefore it has historically been difficult to distinguish the contributing sources of genetic variation when only considering data from a single point in time. However, modern advances in high-throughput sequencing technologies have revolutionized large scale genetic analysis in many ways. In particular, new technologies make it possible to sample the rich complexity of genetic populations densely in time. Thus, the increased accessibility of time-serial genomic data allows for deeper insight into the evolutionary mechanics of populations.

Unlike random sources of variation such as genetic drift, natural selection captures the nonrandom process of how populations adapt to their environment and achieve reproductive success. In particular, the evolutionary impact of natural selection can be seen in a wide spectrum of public health concerns, such as human pathogens, like HIV and influenza, which evolve and adapt rapidly (Illingworth et al., 2012) and tumor cell growth in cancer (Sproufiske et al., 2012). Furthermore, understanding the role selection plays in how populations adapt to environmental influences is central to anticipating the next emergent infectious disease (HIV, West Nile, SARS, ebola), keeping the flu vaccine one step ahead of the influenza virus, prolonging treatment effectiveness (HIV, HBV, HCV, malaria), and personalized treatment of cancer. As a result, in addition to being able to isolate selection from other forces, scientists are increasingly interested in the ability to infer which loci are undergoing selection (detection) and quantifying the selective force (estimation). A complicating factor in these efforts is the fact that the efficacy of selection is affected by the size of the population. For instance, genetic drift in small populations is more likely to overwhelm the selective force, whereas the same selective force will have a more pronounced impact relative to genetic drift within a larger population. Further complicating the matter, natural populations commonly fluctuate in size over time. Viruses, for instance, may experience several shifts in population size as they attempt to replicate while the host defends itself. Under the assumptions of neutral theory, which claims that the majority of genetic variation is a neutral result of genetic drift (Kimura, 1991), evidence for or against selection can be supported by leveraging the added power of temporal changes in allele frequencies in distinguishing selection from genetic drift.
1.2 Selection and Theoretical Population Genetics

Hypothesis tests for selection reject the null hypothesis of neutral evolution and conclude selection if other assumption violations are unlikely or excluded. However, even the simplest neutral models are numerically intractable for all but the smallest populations and shortest evolutionary times. As a result, mathematical approximations that assume large population sizes are ubiquitous. In this section, we introduce the models forming the basis of the work in this dissertation.

1.2.1 Wright-Fisher Model

The Wright-Fisher (WF) idealized population model is the null model underlying numerous neutrality tests and selection detection methods in population genetics. The WF model assumes a single panmictic population, constant in size $N$, with discrete, synchronized generations subject to genetic drift alone (i.e. no mutation, migration, recombination or selection). One of the simplest models in population genetics, it completely specifies the exact distribution of allele frequencies in the next generation given the frequency in the current generation,

\[ X(t+1) \mid X(t) \sim \text{Binomial} \left( N, \frac{X(t)}{N} \right), \tag{1.1} \]

where $X(t)$ is the number of copies (frequency) of the allele of interest in the $t$th generation. Thus, changes in allele frequency, or allele trajectories through time, are modeled as a Markov process with transition probabilities $p_{mn} = \Pr[X(t+1) = n \mid X(t) = m]$ equal to binomial probabilities (Ewens, 2004).

The model is easily generalized to handle selection or mutation. This thesis is concerned with selection over short time periods, when it is reasonable to neglect mutation but neutrality may not hold. Suppose the allele of interest has fitness $1 + s$ relative to other allele(s) at the locus, then

\[ X(t+1) \mid X(t) \sim \text{Binomial} \left( N, \frac{(1 + s)X(t)}{N + sX(t)} \right), \]

so long as $s > -1$. If the selection coefficient $s > 0$, then the allele of interest has a selective advantage over the other allele(s). If $s < 0$, then selection acts against the allele of interest.
Real populations, of course, are not panmictic, nor do most synchronize reproduction. Fortunately, many relaxations of the WF model can merely replace the true number of individuals \( N \) in the real population with a (usually smaller) variance effective population size \( N_e \), the size of an idealized Wright-Fisher population that would accumulate the same amount of variance in the allele frequencies in time. While this substitution of \( N_e \) is convenient, it does not solve all problems with the WF model. In particular, the mathematical computations necessary to exactly evaluate the allele frequency distribution multiple generations into the future quickly become intractable. A solution is to approximate the Markov chain with a continuous time diffusion approximation (Wright, 1945; Kimura, 1957).

1.2.2 Wright-Fisher Diffusion

Consider time, previously measured in generations \( t \), rescaled in units of \( N_e \) as \( \tau \) and redefine \( X(\tau) \) as the relative allele frequency, which I may continue to refer to as “frequency” as is popular in genetics, on the interval \([0,1]\). Define \( f(x \mid p, \tau) = \Pr[X(\tau) = x(\tau) \mid X(0) = p] \) as the probability of the allele frequency at \( \tau \) time units after it started at \( p \). Assume scaled time \( \tau \) and allele frequency \( X(\tau) \) are now continuous, which is reasonable as \( N_e \to \infty \). Then the transition probabilities satisfy the Kolmogorov backward equation (Kimura, 1957)

\[
\frac{\partial}{\partial \tau} f(x \mid p, \tau) = a(p) \frac{\partial}{\partial p} f(x \mid p, \tau) + \frac{1}{2} b(p) \frac{\partial^2}{\partial p^2} f(x \mid p, \tau),
\]

where

\[
a(p) = 0 \quad (1.3)
\]

\[
b(p) = p(1 - p)
\]

are commonly referred to as the infinitesimal drift and diffusion (variance) coefficients (Ewens, 2004). In the default diffusion approximation to the neutral WF model (WFD), there is no drift \( a(p) = 0 \), but WF generalizations allowing mutation, migration or selection are possible by considering different drift terms. For example, if the allele is subject to selection with selection coefficient \( s \) in a haploid population, then

\[
a(p) = N_e s p(1 - p). \quad (1.4)
\]
The Kolmogorov forward equation

\[
\frac{\partial}{\partial \tau} f(x \mid p, \tau) = -\frac{\partial}{\partial x} \left[ a(x) f(x \mid p, \tau) \right] + \frac{1}{2} \frac{\partial^2}{\partial x^2} \left[ x(1 - x) f(x \mid p, \tau) \right] \tag{1.5}
\]

has also been used for various purposes (Wright, 1945; Kimura, 1955; Zhao et al., 2013).

### 1.2.3 Hidden Markov Model for Observational Data

In practice, the true allele frequencies \( X(t) \) are not observed. Instead, experimentalists observe
sample mutant allele counts \( Y(t) \) at specific generations \( t \in \{ t_0, t_1, \ldots, t_T \} \). Under typical sampling
assumptions, the result is the Wright-Fisher Hidden Markov model (WF-HMM), where \( Y(t) \) is
observed at discrete times, and \( X(t) \) is hidden. Indexing the sampling times by \( j = 0, \cdots, T \),
setting \( Y_j = Y(t_j) \) and \( X_j = X(t_j) \) and the corresponding sample sizes \( n_j \), the emission probabilities
\( \Pr[Y_j \mid X_j] \) are typically assumed to be

\[
Y_j \mid X_j \sim \text{Binomial} \left( n_j, X_j \right). \tag{1.6}
\]

As \( N_e \) increases, the WF-HMM becomes intractable and the WF diffusion approximation leads
to a hidden continuous time Markov chain (WFD-HMM). In rescaled time, \( X_j = X(\tau_j) \) are con-
tinuous, and the transition probabilities \( \Pr[X_j \mid \Delta \tau_j, X_{j-1}] \) are solutions to Kolmogorov backward
equation (1.2). Numerical solutions to this system are equally challenging, though there have been
some attempts (Bollback et al., 2008; Williamson and Slatkin, 1999; Anderson et al., 2000).

### 1.2.4 Gaussian Approximation to the WF Diffusion

The WF neutral diffusion can be written as the stochastic differential equation (SDE)

\[
dX(\tau) = \sqrt{X(\tau)[1 - X(\tau)]}dW(\tau)
\]

for Weiner process \( W(\tau) \). The SDE is, by the Euler-Maruyama discretization (Higham, 2001), itself
the limit of discrete time process

\[
\Delta X_n = X_{n+1} - X_n = \sqrt{\Delta \tau X_n(1 - X_n)} \epsilon_n \tag{1.7}
\]
where $\Delta \tau$ is the time increment between subsequent time points and $\epsilon_n \sim \mathcal{N}(0,1)$. Since $\Delta \tau$ is measured in $N_e$ units,

$$
\Delta X_n \sim \mathcal{N}\left(0, \frac{\Delta t X_n(1 - X_n)}{N_e}\right)
$$

(1.8)

if the lapse in generations $\Delta t$ is small. Note, if time lapses $\Delta \tau_j$ are small and sample sizes $n_j$ are large, then the WFD-HMM can be solved using the Kalman filter (Harvey, 1990).

1.3 Estimating/Testing Selection

Controversy over the ubiquity of neutral theory (Ewens, 2004) and interest in identifying the specific variants responsible for adaptation (Stephan, 2016) have lead to many tests of neutrality, some of which use temporal data. While neutral alleles can change in frequency because of genetic drift, it would be unusual to see a consistently increasing or decreasing allele frequency without the assistance of selection. As a result, several methods have begun to leverage the added power of the emerging temporal datasets to detect selection events (Terhorst et al., 2015). It was already possible to estimate major genetic parameters such as effective population size $N_e$ from small temporal datasets (Williamson and Slatkin, 1999; Anderson et al., 2000). With the arrival of larger datasets, it became possible to estimate selection jointly with population size (Bollback et al., 2008), the age of the allele (Malaspinas et al., 2012), while accounting for spatial structure (Mathieson and McVean, 2013) and genetic linkage (Illingworth and Mustonen, 2011; Illingworth et al., 2012).

1.3.1 Bollback et al. (2008)

Bollback et al. (2008) set out to use the WFD-HMM to estimate $N_e$ and $s$ given time series allele frequency data. It is applicable when selection $s \sim o(1/N_e)$ is weak, as $N_e \to \infty$ and assumes initial allele frequency $x_0 \sim \text{Unif}(0,1)$ (Malaspinas, 2016; Hui and Burt, 2015). Define

$$
f_{X_j} = \Pr[Y_j = y_j, \cdots, Y_0 = y_0, X_j = x_j].
$$

(1.9)

Let $\Delta \tau_j = \tau_j - \tau_{j-1}$ be the rescaled elapsed time between two successive sampling timepoints. Under the assumption of conditional independence of the $Y_j$ given $X_j$, in conjunction with the
underlying Markov process characterized by Eq. (1.2), Eq. (1.9) can be written recursively as

\[ f_{X_j} = \Pr[Y_j = y_j \mid X_j = x_j] \int_0^1 f_{X_{j-1}}(x_j \mid x_{j-1}, \tau_j) \, dx_{j-1}. \]  

(1.10)

Thus, the full model likelihood is defined by

\[ \Pr[Y_T = y_T, \ldots, Y_0 = y_0] = \int_0^1 f_{X_T} \, dx_T = \int_0^1 \Pr[Y_T = y_T \mid X_T = x_T] \times \left( \int_0^1 f_{X_{T-1}}(x_T \mid x_{T-1}, \tau_T) \, dx_{T-1} \right) \, dx(\tau_T). \]  

(1.11)

The likelihood Eq. (1.11) with drift Eq. (1.4) defines the Bollback-York-Nielsen model (Bollback et al., 2008) I will denote as BYN(s). In Bollback et al. (2008), the BYN(s) likelihood is evaluated in a two-step process. Solutions to the partial differential equation Eq. (1.2) are numerically calculated using the Crank-Nicolson central finite differencing method (Crank and Nicolson, 1947). Secondly, the integral in equation Eq. (1.10) is numerically approximated using the quadrature midpoint rule. Both steps utilize a grid of all timepoints and possible values for \( X_j \). Finally, the process is repeated on a fixed grid of parameters \( N_e s \) and \( N_e \) to ensure a smooth likelihood surface for visual maximization (Bollback et al., 2008). Inference about the strength of selection \( s \) is via confidence intervals, but their reliance on the asymptotic properties of maximum likelihood estimators can lead to high false positive rates (Feder et al., 2014).

### 1.3.2 Frequency Increment Test

The Frequency Increment Test (FIT) uses the Gaussian approximation to derive an unbiased test of neutrality (Feder et al., 2014). Assume the elapsed times \( \Delta \tau_i = \tau_i - \tau_{i-1} \) between two timepoints \( \tau_i \) and \( \tau_{i-1} \) for \( i = 1, \ldots, T \) are small in comparison to the variance population size \( N_i \) for the \( i \)th interval. Defining the standardized difference in allele frequencies as \( \Delta X_i = \frac{X_i - X_{i-1}}{\sqrt{X_{i-1}(1-X_{i-1})}} \), the standardized frequency increment is normal under the assumption of neutrality,

\[ Y_i = \frac{\Delta X_i}{\sqrt{2\Delta \tau_i}} \sim \mathcal{N} \left( 0, \frac{1}{N_i} \right). \]  

(1.12)
Assuming $N_i = N_e$ for time points $i$, the frequency increment test is constructed using the sample mean $\bar{Y} = \frac{1}{T} \sum_{i=1}^{T} Y_i$ and sample variance $S^2 = \frac{1}{T-1} \sum_{i=1}^{T} (Y_i - \bar{Y})^2$ to form the corresponding Student’s $t$ statistic,

$$t_{FI_i} = \frac{\bar{Y}}{\sqrt{\frac{S^2}{T}}}$$

and rejecting neutrality for sufficiently large values of $|\bar{Y}|$.

### 1.3.3 Frequency Increment Test with Reference Loci

To handle the reality that natural populations often change in size, Nishino (2013) extended FIT to detect selection at a single diallelic locus in the presence of population size variation, $i.e.$ $N_i \neq N_e$ for all time points $i$. Nishino (2013) demonstrated FITR is more powerful than FIT when population sizes are fluctuating.

In practice, $N_i$ is an unknown nuisance parameter and while several methods have attempted to estimate this parameter (Williamson and Slatkin, 1999; Anderson et al., 2000; Bollback et al., 2008; Malaspinas et al., 2012), Nishino (2013) uses $R$ independently and neutrally evolving reference loci to provide information about $N_i$. Define $Y_{ri}$ to be the standardized frequency increment of the $r$th reference locus at time $\tau_i$, where $r = 0, 1, \cdots, R$ with $r = 0$ the current locus to test,

$$Y_{ri} = \frac{\Delta X_{ri}}{\sqrt{\frac{\Delta X_{ri}^2}{2N_i}}} \sim \mathcal{N}(0, 1).$$

The FITR test statistic

$$t_{FITR_i} = \frac{Y_{0i}}{\sqrt{\frac{1}{R} \sum_{r=1}^{R} Y_{ri}^2}} = \frac{\Delta X_{0i}}{\sqrt{\frac{1}{R} \sum_{r=1}^{R} X_{ri}^2}}$$

follows a Student’s $t$ distribution with $R$ degrees of freedom.

### 1.4 Outline of dissertation.

The remainder of this dissertation aims to develop methods to detect sites under the influence of selection. Chapter 2 introduces a porcine reproductive and respiratory syndrome virus (PRRSV) genetic time series dataset, provided by the S. Carpenter lab and the PRRS Host Genetics Consortium, and the challenges encountered when using it to detect selection. Analysis
of molecular variance (AMOVA) evaluates the amount of genetic variation attributed to various sources of population structure, an indirect indication of selection if large genetic change associates with structure thought to induce selective pressures. The Cochran-Mantel-Haenszel (CMH) can test whether genetic change associates with time, while accounting for the population structure. However, despite evidence of significant temporal trends, this method cannot determine whether the trend was encouraged by selection or merely the less interesting consequence of pure genetic drift. This chapter ends with a conjecture, supported through extensive empirical simulation, about the temporal change in allele frequency that can be used as the basis for a probabilistic neutrality test. In chapter 3 we address the issue of the non-normality of relative allele frequency increments in an effort to extend the utility of the FITR. After first applying a normalizing transformation to the discrete, relative allele frequency increments, the assumptions of the Gaussian approximation to the WF-HMM are satisfied. Several normalizing transformations are applied to simulated data, and the results show improvement in the overall skewness and excess kurtosis across all proposed model transformations. We provide evidence that FITR performance is affected by an interaction between the initial allele frequency and effective population size. In regards to test performance, we consistently see better conservation of the FITR type-1 error rate after transformation, although there is no uniformly best choice of transformation across the varying simulation scenarios. Chapter 4 summarizes the key findings of the analyses from chapter 2 and chapter 3 and identifies potential areas of future research development. Applications of alternative normalizing transformations, linkage disequilibrium and alternative approximations to the WF-HMM are briefly discussed. Lastly, we conclude by broadly connecting the implications of this research to a numerous problems in virology.
CHAPTER 2. THE CHALLENGE OF DETECTING SELECTION IN PRRSV

2.1 Introduction

Porcine reproductive and respiratory syndrome virus (PRRSV) is a 15kb positive-stranded enveloped RNA virus that began causing respiratory symptoms and abortions in pig herds in the late 1980’s (Loula, 1991; Chang et al., 2002; Hopper et al., 1992; Christianson et al., 1992). PRRSV is a highly variable virus, characterized by extensive genetic heterogeneity among field isolates and repeated evolution of new variation (Brar et al., 2014), yet little is known about what parts of the virus respond to host selective pressures.

There is an inbuilt difficulty in testing for selection in PRRSV genes. Abundant overlapping reading frames render traditional statistical tests of selection based on dN/dS ratios (Li, 1993; Yang, 1998; Yang and Nielsen, 2000) inappropriate for the envelope protein-encoding ORF2–6. Nevertheless, there is indirect evidence of selection, particularly in the ORF5-encoded envelope protein GP5, which experiences rapid evolution in the ectodomain (Meng et al., 1995). A good review of recent evidence for selection in PRRSV can be found in Evans et al. (2017).

The PRRS Host Genetics Consortium (PHGC), funded by the US National Pork Board, conducted a multi-year, multi-phase study to assess pig resistance and susceptibility to primary PRRS viral infection (Lunney and Chen, 2010). During each of the first four PHGC trials, approximately 200 pigs were infected with in vitro-passaged PRRS virus derived from the reference genome (NVSL 97-7895). Samples were taken from the pigs at several time points post infection, and the virus in some samples was sequenced. The resulting time series of virus sequence data, which we often reduce to the observed counts of nucleotide alleles, A, C, G, or T, at particular genomic sites, may reveal selection when there is an unusual change in the relative frequency of an allele over time. A
particular strength and uniqueness of these data are the replicate pigs, which is useful if selection acts equivalently across pigs.

Our goal is to check for evidence of selection in the PRRSV genome in this subset of pigs. Specifically, we introduce and characterize the virus sequence data, describe the use of the Cochran-Mantel-Haenszel (CMH) test, or equivalent tests, to detect significantly varying sites, explain why the CMH test is insufficient to identify selected sites, and propose a novel test for detecting when such variation may reflect actual selection at the site. We find no definitive evidence for selection acting on this dataset, in part because of the difficulty of detecting selection from temporal data without knowledge of the effective population size $N_e$ and in part because of the characteristics of this particular dataset.

2.2 Methods

2.2.1 Samples, sequences, alignments and SNVs

In the PHGC trials, blood samples and weight measurements were taken at various sampling times during a 42-day observation period. Total virus load was measured at all time points
(Fig. 2.1), but we analyze two small sequencing datasets obtained from the blood samples of a small number of Large White and Landrace commercial crossbred pigs. One study, the “virus study”, isolated viral RNA from seven pigs in order to understand evolution of the virus during early infection (Evans et al., 2017). The other study, the “host study”, isolated mRNA from 16 pigs in order to understand host gene expression changes in response to infection (Schroyen et al., 2016). Among the host mRNA, we found considerable amounts of PRRSV mRNA, which we consider our second dataset.

For the virus study, serum samples were collected from five pigs involved in PHGC trials one and three (Boddicker et al., 2012). Two maintained high levels of viremia throughout 35 dpi (prolonged), and three initially cleared the virus only to experience rebound viremia by 41 dpi (rebound). Viral RNA was isolated from the NVSL97-7895-derived inoculum, day 7 sera from all pigs, and late day sera from prolonged and rebound pigs as described in Evans et al. (2017). Briefly, viral RNA was isolated from sera, reverse transcribed to cDNA, and then ORF2–6 and nsp2 were amplified using PRRSV-specific primers. Individual PCR products were cloned and several positive clones were selected for Sanger sequencing. The sequences from each genetic region were separately aligned using Muscle version 3.8.31 (Edgar, 2004). Three insertions (each occurring in six or fewer sequences) in nsp2 were removed for the site-by-site tests of selection, but retained for the studies of genetic structure. The sampled time points and number of clones sequenced for each genetic region of the pigs are given in Table 2.1.

In the host study, whole blood samples were collected from sixteen pigs sampled from the third PHGC trial to study host gene transcription. Sampling time points for each pig were zero (prior to infection), four, seven, ten and fourteen days post inoculation (dpi). Details of the library preparation and sequencing are in Schroyen et al. (2016). Briefly, total RNA was isolated, globin reduce by RNase H, the library prepared by the TrueSeq™ library kit (Illumina, Inc., San Dieo, CA, USA), followed by production of 50bp paired end reads on the Illumina HiSeq in eight lanes. To isolate the PRRSV sequences from those of the pig host, sequences were aligned to the pig genome assembly (GenBank: GCA_000003025.4) and PRRSV NVSL97-7895 reference
Table 2.1: Number of PRRSV clones sequenced in inoculum and from samples taken at specified days post infection (dpi) in selected pigs. The source PHGC trial and virological outcome (see text) of each pig is also noted.

<table>
<thead>
<tr>
<th>Pig ID</th>
<th>PHGC Trial</th>
<th>Virological Outcome</th>
<th>dpi</th>
<th>Number of clones</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Inoculum</td>
<td>ORF2–6</td>
<td>nsp2</td>
<td></td>
</tr>
<tr>
<td>C-1165</td>
<td>1</td>
<td>Cleared</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>C-3187</td>
<td>3</td>
<td>Cleared</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>P-1134</td>
<td>1</td>
<td>Prolonged</td>
<td>7</td>
<td>6, 20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>28</td>
<td>20, 33</td>
</tr>
<tr>
<td>P-3161</td>
<td>3</td>
<td>Prolonged</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>28</td>
<td>25, 24</td>
</tr>
<tr>
<td>R-1113</td>
<td>1</td>
<td>Rebound</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>35</td>
<td>24, 25</td>
</tr>
<tr>
<td>R-3068</td>
<td>3</td>
<td>Rebound</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>35</td>
<td>26, 30</td>
</tr>
<tr>
<td>R-3197</td>
<td>3</td>
<td>Rebound</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>41</td>
<td>23, 32</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PHGC</th>
<th>Virological</th>
<th>Number of clones</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ORF2–6</td>
<td>nsp2</td>
</tr>
<tr>
<td></td>
<td>176</td>
<td>204</td>
</tr>
</tbody>
</table>

genome (GenBank: AY545985.1) using bwa-mem version 0.7.5 (Li and Durbin, 2010) with default parameters. Ambiguous sequences that mapped to both the pig genome and the reference genome were removed. ORF2–6 were found to be much more highly sequenced than other PRRSV genomic regions, so we only retained reads aligning to ORF2–6, where coverage was high. Average coverage for each ORF of 13 pigs is shown in Table 2.2. Three pigs were removed for lacking sufficient data for at least one of the five sampling timepoints.

In all, we have three aligned datasets: ORF2–6 clonal sequences from the virus study, nsp2 clonal sequences from the virus study, and ORF2–6 RNA-seq from the host study. We consider each possible site in an alignment a Single Nucleotide Variant (SNV) if it shows any variation. In the RNA-seq data, such sites may not be true SNVs because of elevated sequencing error rates, but sequencing errors should not display systematic trends across time or pigs and will be handled as noise. Within each dataset, we record the frequency of each observed nucleotide within each site/pig/time point combination. For each site in each dataset, we also record the major allele that
Table 2.2: Pig IDs and average number of NGS reads spanning various genes in ORF2–6 in the host study.

<table>
<thead>
<tr>
<th>Pig ID</th>
<th>Overall</th>
<th>E</th>
<th>GP2</th>
<th>GP3</th>
<th>GP4</th>
<th>GP5</th>
<th>GP5a</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>03031</td>
<td>1593</td>
<td>887</td>
<td>1430</td>
<td>2952</td>
<td>4551</td>
<td>8906</td>
<td>7102</td>
<td>15689</td>
</tr>
<tr>
<td>03068</td>
<td>2091</td>
<td>1474</td>
<td>2328</td>
<td>4266</td>
<td>6366</td>
<td>11763</td>
<td>9895</td>
<td>19271</td>
</tr>
<tr>
<td>03074</td>
<td>1972</td>
<td>1687</td>
<td>2422</td>
<td>4460</td>
<td>6402</td>
<td>11041</td>
<td>9301</td>
<td>17041</td>
</tr>
<tr>
<td>03089</td>
<td>4449</td>
<td>4929</td>
<td>6785</td>
<td>11549</td>
<td>16111</td>
<td>23979</td>
<td>21154</td>
<td>34640</td>
</tr>
<tr>
<td>03094</td>
<td>744</td>
<td>893</td>
<td>1207</td>
<td>1910</td>
<td>2571</td>
<td>4090</td>
<td>3524</td>
<td>5771</td>
</tr>
<tr>
<td>03101</td>
<td>3048</td>
<td>4457</td>
<td>5759</td>
<td>8547</td>
<td>10919</td>
<td>17495</td>
<td>15409</td>
<td>20857</td>
</tr>
<tr>
<td>03112</td>
<td>1198</td>
<td>1052</td>
<td>1545</td>
<td>2906</td>
<td>3824</td>
<td>6727</td>
<td>5487</td>
<td>10079</td>
</tr>
<tr>
<td>03113</td>
<td>832</td>
<td>300</td>
<td>541</td>
<td>1395</td>
<td>2216</td>
<td>4582</td>
<td>3739</td>
<td>8886</td>
</tr>
<tr>
<td>03148</td>
<td>2165</td>
<td>855</td>
<td>1587</td>
<td>3818</td>
<td>5868</td>
<td>12068</td>
<td>9813</td>
<td>22203</td>
</tr>
<tr>
<td>03159</td>
<td>1631</td>
<td>732</td>
<td>1352</td>
<td>3017</td>
<td>4479</td>
<td>9364</td>
<td>7426</td>
<td>16133</td>
</tr>
<tr>
<td>03170</td>
<td>812</td>
<td>256</td>
<td>526</td>
<td>1334</td>
<td>2186</td>
<td>4502</td>
<td>3643</td>
<td>8623</td>
</tr>
<tr>
<td>03174</td>
<td>813</td>
<td>891</td>
<td>1222</td>
<td>2086</td>
<td>2926</td>
<td>4501</td>
<td>4000</td>
<td>6239</td>
</tr>
<tr>
<td>03192</td>
<td>808</td>
<td>976</td>
<td>1326</td>
<td>2094</td>
<td>2829</td>
<td>4448</td>
<td>3831</td>
<td>6191</td>
</tr>
</tbody>
</table>

is most common across pigs and times, and for some analyses we aggregate the counts of all minor alleles into the combined minor allele frequency.

2.2.2 AMOVA

The sampled virus population is highly structured. In the virus study, virus genotypes were sampled from multiple days in multiple pigs stratified by clinical disease outcome in two experimental trials (Table 2.1). To examine whether the population structure produced genetic structure in the sampled data, we used Analysis of Molecular Variance (AMOVA) (Excoffier et al., 1992), which requires a distance between all pairs of sampled sequences. We use a continuous time Markov chain model (CTMC) to compute evolutionary distances (Ewens, 2004). CTMC model selection is done with jModelTest, version 2.1.6 (Darriba et al., 2012). We run jModelTest, with Phyml version 3 (Guindon et al., 2010), to examine 11 substitution schemes, with and without equal base frequencies (+F), invariant sites (+I), and gamma distributed rate variation (+G) in four categories, with free parameters optimized on the BIONJ tree. The TVM+G model was selected by
BIC for ORF2–6, while TPM2uf+G was selected by BIC for nsp2. Final parameter values of these selected models were estimated using Phylm from 10 random initial trees, estimating equilibrium base frequencies from the observed empirical frequencies, optimizing all other parameters (topology, branch lengths, and rate parameters), and using the best of NNI and SPR topology search. The same likelihood maximum was found in at least 3 of the 10 random initial trees. Parameter values estimated from the initialization resulting in highest likelihood were used in subsequent calculations. Finally, Tree-puzzle (Schmidt et al., 2002) was used to compute pairwise distances between all sequences using these models and estimated parameters.

Custom perl scripts were used to prepare data files for AMOVA as implemented in the R package ade4 (Dray and Dufour, 2007). Specifically, we provided a sample, structure, and distance file to amova() and tested variance components with 10,000 random permutations using randtest().

### 2.2.3 Testing association of alleles and time

To determine whether there is a pattern of variation over time, we use the Cochran-Mantel-Haenszel (CMH) test (Cochran, 1954; Mantel and Haenszel, 1959) or equivalently, the conditional logistic regression score test when the response is binary (Day and Byar, 1979). These tests identify associations between two categorical variables while controlling for other covariates. In particular, we use them to detect an association between SNVs and time, while aggregating information across pigs.

In the virus study, we aggregate all minor alleles into a combined minor allele and then test for association of the major allele and dpi, assuming a constant effect of each distinct dpi (0, 7, 28, 35, and 41) across pigs, while stratifying on the source pig/experimental trial. Since the five pigs of the virus study with multiple time points were inoculated with the same virus pool, we randomly and uniformly partition the available inoculum sequences and assign them to the five pigs. We use the clogit() function in R version 3.4.4 (2018-03-15) package survival to test whether dpi has an effect using the score test.
For the host study, we used time points 4, 7, 10 and 14 dpi, discarding the 0 dpi sample because it was taken before infection and hence should contain no PRRSV sequence. We obtained a p-value from the R function `mantelhaen.test` with a two-sided alternative hypothesis. In this case, we retained all four alleles, A, C, G, and T, in the test.

Holm’s correction (Holm, 1979) was used to maintain the familywise error rate below 0.05 given the number of tests (total number of SNVs) within each of the two datasets.

### 2.2.4 Hellams Test

We state without proof the following theorem.

**Theorem 1** (Hellams conjecture). Let $X_t$ be the relative frequency of an allele after $t$ generations of the Wright-Fisher Markov chain or the Wright-Fisher diffusion approximation with appropriate scaling (see Chapter 1). Then,

\[
P(X_t \geq x_0) \leq x_0 \text{ for } x_0 > 0.5, \text{ and} \]
\[
P(X_t \leq x_0) \leq x_0 \text{ for } x_0 < 0.5,
\]

where $x_0$ is the initial relative allele frequency at generation 0. At $x_0 = 0.5$, the inequality becomes an equality.

We provide some empirical justification of this conjecture in the results, and we have consulted two probabilists for a proof, but there is no proof yet. A compelling fact that seems to argue strongly for the result is that the probability of fixation ($\lim_{t \to \infty} X_t = 1$) for an allele starting with initial allele frequency $x_0$ is precisely $x_0$. This fact is true for both the discrete Markov chain and continuous diffusion approximation. It seems unlikely that an allele starting with $x_0 \geq 0.5$ will have more than $x_0$ probability of exceeding its starting value.

The beauty of this conjecture is the fact that no knowledge of the often unknown effective population size $N_e$ is required to apply it. To develop a test for selection, we consider $x_0$ an unknown parameter, $X_t$ a random variable, and define $Z = I(X_t \geq x_0)$. Given $x_0 > 0.5$, no selection and the theorem, we know $P(X_t \geq x_0) \leq x_0$. When there is positive selection for the
allele, then selection coefficient \( s > 0 \) and we expect \( X_t \geq x_0 \) more often than under neutral theory. So, we may test

\[
H_0 : E[Z] \leq x_0
\]

against \( H_1 : E[Z] > x_0 \) to detect positive selection. The test just discussed is intended for application to the major allele at a locus, where \( x_0 > 0.5 \). Clearly it has no power to detect an unusual increase when \( x_0 < 0.5 \), for then we expect \( \Pr(X_t > x_0) \geq 1 - x_0 \). Thus, the test is not useful for detecting positive selection for a newly emerged allele nor negative selection against a major allele.

We do not observe \( X_t \), but rather conditionally independent random variables

\[
Y_0 \sim \text{Bin}(n_0, x_0) \quad \text{and} \quad Y_1 | X_t \sim \text{Bin}(n_t, X_t).
\]

Specifically, we observe the pairs \( (Y_{i0}, Y_{it}) \) for \( 1 \leq i \leq n \). Here, \( n \) may be separate experimental units assumed to be evolving under the same selection pressure. Or \( n \) may be the number of time points after the first in time series data. The sequential increments \( X_{it} - X_{i0} \) are conditionally independent given the Markovian nature of the process. In our PRRSV data, \( n \) is a combination of both. Let \( T_i = I \left( \frac{Y_{it}}{n_{it}} \geq \frac{Y_{i0}}{n_{i0}} \right) \). It is an estimate of \( P(X_t \geq x_0) \) as is \( \hat{x}_{i0} = \frac{1}{n_{i0}} Y_{i0} \) under the point null. Neglecting pig effects, the \( T_i \) are independent but not identically distributed Bernoulli random variables. Therefore the sum \( T = \sum_{i=1}^{n} T_i \) has a Poisson binomial distribution with probabilities of success estimated by \( \hat{x}_{i0} \). We use the \texttt{R poibin} package to compute \( p \)-values for the one-sided test \( H_0 : s > 0 \), equivalent to Eq. \( (2.1) \). We may also, of course, test the other one-sided null \( H_0 : s < 0 \) by considering an increase in the other allele at a biallelic site.

### 2.3 Results

#### 2.3.1 Genetic diversity

Figure 2.2 presents a visualization of the nucleotide variation present in the virus dataset alignments relative to the consensus sequence in the inoculum for ORF2–6 and \( nsp2 \). There is sporadic variation throughout both genes, but there are some sites that appear to change completely or almost completely \textit{in vivo}. In some cases, the change is observed in all pigs, especially at the late
Figure 2.2: Alignments show the ORF2-6 or nsp2 consensus sequences in the inoculum at the top and mutations away from consensus below. Sequences are arranged vertically by pig, then by disease stage (dpi) within pigs. Colors in the plot indicate nucleotide: A=green, C=blue, G=purple, T=red, and deletion=gray. The legend on the right indicates the pig and stage of each block of sequences. The segregating sites with at least one mutation, including possible deletion, are indicated in orange along the horizontal axis at the bottom of the plot.
time points. In others, the change is only in one pig or a few pigs. These sites seem the most likely subjects of selection, but the sweeping allele often arises from preexisting variation in the inoculum and could also have spread via genetic drift, although less and less plausible as more and more pigs experience it.

We quantify the total diversity within pigs and divergence between pigs as a function of time. To quantify diversity/divergence, we computed the pairwise distances between sequences using the methods described for AMOVA. Then, the diversity within a pig at a particular time is the average pairwise distance of sequences sampled from that pig at that dpi. The divergence from the inoculum is the average pairwise distance between sequences sampled from that pig/dpi with the inoculum sequences, and the divergence between pigs is the average pairwise distance between all pairs of sequences from distinct pigs at the same dpi. Figures 2.3a and 2.3b show that divergence and between-pig diversity are generally increasing in time, which is to be expected as the population evolves away from the inoculum. However, especially in ORF2–6, within-pig diversity remains roughly the same as the original inoculum diversity. Overall, there are extremely low amounts of variation.

To look for indirect evidence of selection, we tested whether the known epitopes, parts of the protein targeted by the immune response, in the ORF2–6 genes were unusually divergent compared to the non-epitope regions. We computed the average entropy (across sites) in epitope and non-epitope regions for each pig and day. We then fit a linear model to entropy and found that epitopes are significantly more variable (p-value 0.030). There was no significant association with dpi or any other measure of disease stage.

### 2.3.2 Genetic structure

To verify that the low level genetic variation reflects biological variation rather than technical variation, we use AMOVA, which requires amplicon sequences and can only be applied to the virus study data. Specifically, we expect the population structure, reflecting the experimental design, should induce a genetic structure within ORF2–6 and nsp2. The population structure in this study
Figure 2.3: We plot within-pig diversity (average pairwise distance within pigs) in red, temporal divergence (average pairwise distance to inoculum sequences) in black, and population diversity or between-pig diversity (average pairwise distance between pigs) in blue. Each plotting point represents data from one pig/one time point. Legend in left plot is for color. Legend in right plot is for symbol. Both legends apply to both plots.

Table 2.3: Genetic structure in ORF2–6 and nsp2 determined by AMOVA.

<table>
<thead>
<tr>
<th>Hierarchical Level</th>
<th>ORF2–6 Variance</th>
<th>p-value</th>
<th>F Statistic</th>
<th>nsP2 Variance</th>
<th>p-value</th>
<th>F Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined</td>
<td>4767</td>
<td>&lt;0.001</td>
<td>0.993</td>
<td>5770</td>
<td>&lt;0.001</td>
<td>0.988</td>
</tr>
<tr>
<td>Experimental Trial</td>
<td>3786</td>
<td>0.007</td>
<td>0.789</td>
<td>4661</td>
<td>0.004</td>
<td>0.798</td>
</tr>
<tr>
<td>Pig</td>
<td>870.1</td>
<td>0.007</td>
<td>0.858</td>
<td>982.2</td>
<td>0.003</td>
<td>0.834</td>
</tr>
<tr>
<td>Sample</td>
<td>111.0</td>
<td>&lt;0.001</td>
<td>0.770</td>
<td>126.6</td>
<td>&lt;0.001</td>
<td>0.647</td>
</tr>
</tbody>
</table>

*aSequences were sampled from blood taken at multiple days within each pig, which were in turn grouped in experimental trials, so sample is nested in pig is nested in experimental trial.
*bThe estimated genetic variance component corresponding to the named hierarchical level.
*cThis p-value is from a test on the within day variance component, which is not shown here. dThe F statistic reports the proportion of total variation explained by the given level, while accounting for variance explained by higher levels of the hierarchy.
*eThe sum of genetic variance components from all structural levels of the hierarchy (day, pig, experimental trial) combined.
is hierarchical, in that the samples represent virus from different days within different pigs within different experimental trials (Table 2.1). AMOVA estimates the portion of total genetic variation within each gene region that is explained by each level of the population structure hierarchy: sample day, pig, and experimental trial. Results indicate that all levels of the hierarchy contribute significantly to total variance in both ORF2–6 and \( nsp2 \), as shown by the \( p \)-values reported in Table 2.3. Almost 80% of the total genetic variation in both ORF2–6 and \( nsp2 \) was explained by the experimental trial (Table 2.3). Of the remaining variation not explained by experimental trial, 86% of the variation in ORF2–6 and 83% of the variation in \( nsp2 \) was due to the pig. After genetic variation explained by both experimental trial and pig was removed, the fraction of the remaining variance explained by the sample day was 77% in ORF2–6 and 65% in \( nsp2 \). The component of variance attributable to virological outcome (cleared, prolonged, or rebound) was not distinguishable from zero (\( p \)-value 0.38). In total, the combined population structure hierarchy (experiment, pig, day) explains 99% of the observed genetic variation in both ORF2–6 and \( nsp2 \). This result indicates significant genetic structure within both gene regions that is consistent with the biological population structure and shows that there was little to no exchange of virus between pigs.

### 2.3.3 Testing for association with time

Given the biological relevance of the genetic diversity, we now use conditional logistic regression to detect significant associations between single nucleotide variants (SNVs) and time, or disease stage, across the pigs. All SNV sites in all five pigs with late time points (395 sites in ORF2–6; 489 sites in \( nsp2 \)) were tested for association between the SNV and dpi. After controlling for multiple testing, we identified 13 SNVs in ORF2–6 and 12 SNVs in \( nsp2 \) showing significant association with dpi in the virus study (Table 2.4). Similarly, by the Cochran-Mantel-Haenszel test, which allowed us to retain all four alleles at a site, we found many significant sites in the pig study (Table 2.5).

These tests cannot rule out genetic drift as an explanation for changing relative allele frequencies, but highly selected sites should be among the sites with significant temporal effects. There are two
Table 2.4: Significantly varying SNVs in virus study by conditional logistic regression score test in ORF2–6 (above the middle rule) and nsp2 (below the middle rule). The starred sites are confirmed by the CMH test applied to the RNA-seq data from the host study.

<table>
<thead>
<tr>
<th>Site</th>
<th>NVSL</th>
<th>Inoc</th>
<th>Maj</th>
<th>Min</th>
<th>Location</th>
<th>Amino Acid</th>
<th>p-value</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>*30</td>
<td>G</td>
<td>G</td>
<td>(0.60)</td>
<td>G (0.99)</td>
<td>A (0.01)</td>
<td>E; Gp2</td>
<td>D9N*; L10L</td>
<td>6.69 x 10^{-7}</td>
</tr>
<tr>
<td>134</td>
<td>C</td>
<td>C</td>
<td>(1.00)</td>
<td>C (0.86)</td>
<td>T (0.18)</td>
<td>E; Gp2</td>
<td>G43G; A45V*</td>
<td>4.25 x 10^{-2}</td>
</tr>
<tr>
<td>*159</td>
<td>T</td>
<td>T</td>
<td>(0.60)</td>
<td>T (0.99)</td>
<td>A (0.00)</td>
<td>E; Gp2</td>
<td>L52M*; D53E*</td>
<td>3.41 x 10^{-8}</td>
</tr>
<tr>
<td>*291</td>
<td>G</td>
<td>G</td>
<td>(0.60)</td>
<td>G (0.94)</td>
<td>A (0.08)</td>
<td>Gp2</td>
<td>M97I*</td>
<td>3.63 x 10^{-3}</td>
</tr>
<tr>
<td>711</td>
<td>A</td>
<td>A</td>
<td>(0.95)</td>
<td>A (0.85)</td>
<td>G (0.20)</td>
<td>Gp2; Gp3</td>
<td>I237M*; T30A*</td>
<td>4.29 x 10^{-5}</td>
</tr>
<tr>
<td>*909</td>
<td>C</td>
<td>C</td>
<td>(0.55)</td>
<td>C (0.80)</td>
<td>T (0.11)</td>
<td>Gp3</td>
<td>P96S*</td>
<td>6.51 x 10^{-4}</td>
</tr>
<tr>
<td>*1050</td>
<td>T</td>
<td>C</td>
<td>(0.50)</td>
<td>T (0.99)</td>
<td>C (0.00)</td>
<td>Gp3</td>
<td>L143L</td>
<td>1.08 x 10^{-11}</td>
</tr>
<tr>
<td>*1523</td>
<td>G</td>
<td>G</td>
<td>(1.00)</td>
<td>G (0.74)</td>
<td>A (0.33)</td>
<td>Gp4</td>
<td>G119S*</td>
<td>9.42 x 10^{-5}</td>
</tr>
<tr>
<td>*1538</td>
<td>A</td>
<td>A</td>
<td>(0.60)</td>
<td>A (0.96)</td>
<td>G (0.00)</td>
<td>Gp4</td>
<td>I124V*</td>
<td>1.59 x 10^{-5}</td>
</tr>
<tr>
<td>*1553</td>
<td>G</td>
<td>A</td>
<td>(0.90)</td>
<td>G (0.88)</td>
<td>A (0.01)</td>
<td>Gp4</td>
<td>I129I</td>
<td>0</td>
</tr>
<tr>
<td>1795</td>
<td>C</td>
<td>C</td>
<td>(1.00)</td>
<td>C (0.67)</td>
<td>T (0.43)</td>
<td>Gp5; Gp5a</td>
<td>A27V*; C30C</td>
<td>1.07 x 10^{-6}</td>
</tr>
<tr>
<td>*1838</td>
<td>G</td>
<td>G</td>
<td>(1.00)</td>
<td>G (0.85)</td>
<td>A (0.19)</td>
<td>Gp5; Gp5a</td>
<td>L41L; D45N*</td>
<td>0</td>
</tr>
<tr>
<td>1884</td>
<td>A</td>
<td>A</td>
<td>(1.00)</td>
<td>A (0.85)</td>
<td>G (0.19)</td>
<td>Gp5</td>
<td>K57E*</td>
<td>0</td>
</tr>
</tbody>
</table>

| 857  | T    | T    | (1.00) | T (0.83) | C (0.22) | HV2 | 286 | 8.28 x 10^{-5} | 6 |
| 1159 | G    | G    | (0.54) | G (1.00) | T (0.00) | HV2 | 387 | 0 | 1 |
| 1356 | A    | A    | (0.67) | G (0.62) | A (0.34) | HV2 | 453 | 1.57 x 10^{-2} | 10 |
| 1412 | T    | T    | (1.00) | T (0.77) | C (0.23) | HV2 | 471 | 7.23 x 10^{-6} | 5 |
| 1820 | T    | T    | (0.79) | T (1.00) | C (0.00) | HV2 | 607 | 1.15 x 10^{-4} | 7 |
| 1834 | G    | G    | (0.83) | G (0.99) | A (0.00) | HV2 | 612 | 1.14 x 10^{-2} | 9 |
| 1841 | T    | T    | (1.00) | T (0.88) | C (0.10) | HV2 | 614 | 2.33 x 10^{-2} | 11 |
| 1871 | C    | C    | (0.54) | C (1.00) | T (0.00) | HV2 | 624 | 0 | 1 |
| 1911 | A    | A    | (0.96) | A (0.66) | G (0.38) | HV2 | 638 | 3.43 x 10^{-2} | 12 |
| 2359 | C    | C    | (0.67) | C (1.00) | T (0.00) | HV2 | 787 | 1.09 x 10^{-9} | 4 |
| 2664 | G    | G    | (0.58) | A (0.91) | G (0.02) | TM1 | 889 | 6.05 x 10^{-12} | 3 |
| 2685 | T    | T    | (0.75) | T (0.99) | C (0.00) | TM1 | 896 | 4.9 x 10^{-4} | 8 |

\n
\*Nucleotide site within region ORF2–6 or nsp2.

\n
bNucleotide in NVSL97-7895 reference sequence at this site.

\n
cMajor nucleotide in the inoculum sample (relative frequency in the inoculum).

\n
dMajor nucleotide in vivo, across pigs and time points (relative frequency in vivo).

\n
eMost common minor nucleotide in vivo (relative frequency in last days sampled).

\n
fFor ORF2–6, the affected envelope protein(s); for nsp2, the affected domain.

\n
gAssociated amino acid change at location in respective protein; nonsynonymous changes starred.

\n
hAdjusted p-value for test of no change in SNV relative frequency over time; 0 implies a value less than the precision of the R function.

\n
iRank of p-value within each gene.
ways to build evidence for selection as the cause behind these temporal changes: (1) a pattern of genetic drift should not repeat itself in these two independent sets of pigs, and (2) we expect selected sites to be nonsynonymous under the presumption that selection acts on the encoded proteins. We explore each of these possibilities below.

Since the host study constitutes far more data and has more power, we expect most of the significant sites from the virus study to also appear in the host study if they are truly selected. However, only nine of the 13 significant SNVs found in the virus study are replicated in the host study. Sites 20, 134, 711, and 1884 changed from the major allele to a minor allele in pig 3197, the only pig sampled at 41 dpi in the virus study. Meanwhile, the polymorphism remained stable in the 13 pigs of the host study, which were sampled for only two weeks after infection. Thus, if these sites are among the selected, evidence of selection only appears later in disease. A more plausible explanation is that these sites changed only after a severe bottleneck in the pig 3197, which experienced a “rebound” in virus at 41 dpi. The only other site not replicated in the host study was site 1795, which changed in several of the virus study pigs, but remained stable in the host study pigs. Since the virus study required PRRSV-specific primers, the differences could reflect primer-dependent sampling bias. On the other hand, the host study sampled mRNA, which may be affected by expression biases. The latter explanation suggests a role for site 1795 in virus gene expression.

The second aid in distinguishing genetic drift from selection, is to test whether SNV with significant temporal effects tend to be nonsynonymous. For significantly varying sites in Table 2.4, we considered the proportion of nonsynonymous substitutions among all possible mutations at the site in the inoculum consensus context. The proportion of nonsynonymous SNVs in ORF2–6 (10/14) was not unusual (p-value 1.00), but the proportion (2/8) in nsp2 was unusually low (p-value 0.02). Strong selective pressure against nonsynonymous change could explain the nsp2 results, but this method is also primitive in weighting all substitutions equally when it is well-known they are not created equal (Graur and Li, 2000).
Taking all the evidence together, we conclude there are significant temporal trends in relative allele frequencies, but we cannot directly conclude that selection has been the cause. The starred sites in ORF2–6 of Table 2.4 (no confirmation is possible for sites in nsp2) are plausible objects of selection, because they have been replicated in the host study. The vast numbers of remaining sites detected in ORF2–6 from the host study (Table 2.5) may either move because of genetic drift or selection. The CMH test has not provided a clear answer.

Table 2.5: Significantly varying SNVs in host study by Cochran-Mantel-Haenszel (CMH) test in ORF2–6. This is table has fixed placement and hence may appear out of sequence with other floating Tables. Please see Table 2.4 for an explanation of the columns.

<table>
<thead>
<tr>
<th>Site</th>
<th>NVSL</th>
<th>Inoc</th>
<th>Maj</th>
<th>Min</th>
<th>Location</th>
<th>Amino Acid</th>
<th>p-value</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>C</td>
<td>C (1.00)</td>
<td>C (0.98)</td>
<td>A (0.01)</td>
<td>E; Gp2</td>
<td>L7L; S9Y*</td>
<td>$6.1 \times 10^{-9}$</td>
<td>22</td>
</tr>
<tr>
<td>30</td>
<td>G</td>
<td>G (0.60)</td>
<td>G (0.97)</td>
<td>A (0.01)</td>
<td>E; Gp2</td>
<td>D9N*; L10L</td>
<td>$2.05 \times 10^{-14}$</td>
<td>19</td>
</tr>
<tr>
<td>44</td>
<td>A</td>
<td>A (1.00)</td>
<td>A (0.98)</td>
<td>G (0.01)</td>
<td>E; Gp2</td>
<td>Q13Q; N15S*</td>
<td>$6.32 \times 10^{-3}$</td>
<td>57</td>
</tr>
<tr>
<td>68</td>
<td>A</td>
<td>A (1.00)</td>
<td>A (0.98)</td>
<td>G (0.01)</td>
<td>E; Gp2</td>
<td>E21E; N23S*</td>
<td>$1.21 \times 10^{-8}$</td>
<td>25</td>
</tr>
<tr>
<td>83</td>
<td>T</td>
<td>T (1.00)</td>
<td>T (0.99)</td>
<td>G (0.00)</td>
<td>E; Gp2</td>
<td>I26M*; L28W*</td>
<td>$3.23 \times 10^{-2}$</td>
<td>68</td>
</tr>
<tr>
<td>92</td>
<td>C</td>
<td>C (1.00)</td>
<td>C (0.98)</td>
<td>T (0.01)</td>
<td>E; Gp2</td>
<td>I29I; S31L*</td>
<td>$3.67 \times 10^{-2}$</td>
<td>79</td>
</tr>
<tr>
<td>156</td>
<td>C</td>
<td>C (1.00)</td>
<td>C (0.98)</td>
<td>T (0.01)</td>
<td>E; Gp2</td>
<td>R51**; S52S</td>
<td>$2.07 \times 10^{-3}$</td>
<td>48</td>
</tr>
<tr>
<td>159</td>
<td>T</td>
<td>T (0.60)</td>
<td>T (0.97)</td>
<td>A (0.01)</td>
<td>E; Gp2</td>
<td>L52M*; D53E*</td>
<td>$4.23 \times 10^{-9}$</td>
<td>21</td>
</tr>
<tr>
<td>174</td>
<td>G</td>
<td>G (1.00)</td>
<td>G (0.98)</td>
<td>C (0.01)</td>
<td>E; Gp2</td>
<td>V57L*; R58R</td>
<td>$3.61 \times 10^{-2}$</td>
<td>78</td>
</tr>
<tr>
<td>192</td>
<td>A</td>
<td>A (1.00)</td>
<td>A (0.98)</td>
<td>C (0.01)</td>
<td>E; Gp2</td>
<td>T63P*; L64L</td>
<td>$1.15 \times 10^{-3}$</td>
<td>44</td>
</tr>
<tr>
<td>201</td>
<td>C</td>
<td>C (1.00)</td>
<td>C (0.99)</td>
<td>T (0.01)</td>
<td>E; Gp2</td>
<td>P66S*; T76T</td>
<td>$1.03 \times 10^{-5}$</td>
<td>33</td>
</tr>
<tr>
<td>222</td>
<td>C</td>
<td>C (1.00)</td>
<td>C (0.98)</td>
<td>A (0.01)</td>
<td>E; Gp2</td>
<td>L73I*; S74S</td>
<td>$3.13 \times 10^{-3}$</td>
<td>45</td>
</tr>
<tr>
<td>240</td>
<td>T</td>
<td>T (1.00)</td>
<td>T (0.99)</td>
<td>C (0.01)</td>
<td>Gp2</td>
<td>S80S</td>
<td>$1.16 \times 10^{-7}$</td>
<td>27</td>
</tr>
<tr>
<td>243</td>
<td>G</td>
<td>G (1.00)</td>
<td>G (0.99)</td>
<td>T (0.01)</td>
<td>Gp2</td>
<td>Q81H*</td>
<td>$10 \times 10^{-3}$</td>
<td>63</td>
</tr>
<tr>
<td>249</td>
<td>G</td>
<td>G (1.00)</td>
<td>G (0.99)</td>
<td>C (0.00)</td>
<td>Gp2</td>
<td>Q3H*</td>
<td>$3.62 \times 10^{-3}$</td>
<td>51</td>
</tr>
<tr>
<td>255</td>
<td>T</td>
<td>T (1.00)</td>
<td>T (0.99)</td>
<td>C (0.01)</td>
<td>Gp2</td>
<td>D58D</td>
<td>$7.34 \times 10^{-9}$</td>
<td>24</td>
</tr>
<tr>
<td>261</td>
<td>C</td>
<td>C (1.00)</td>
<td>C (0.99)</td>
<td>T (0.00)</td>
<td>Gp2</td>
<td>P87P</td>
<td>$2.24 \times 10^{-2}$</td>
<td>72</td>
</tr>
<tr>
<td>275</td>
<td>G</td>
<td>G (1.00)</td>
<td>G (0.99)</td>
<td>A (0.00)</td>
<td>Gp2</td>
<td>R92K*</td>
<td>$2.06 \times 10^{-6}$</td>
<td>30</td>
</tr>
<tr>
<td>282</td>
<td>C</td>
<td>C (1.00)</td>
<td>C (0.99)</td>
<td>A (0.00)</td>
<td>Gp2</td>
<td>P94P</td>
<td>$3.92 \times 10^{-8}$</td>
<td>26</td>
</tr>
<tr>
<td>283</td>
<td>C</td>
<td>C (1.00)</td>
<td>C (0.99)</td>
<td>T (0.00)</td>
<td>Gp2</td>
<td>L95L</td>
<td>$5.73 \times 10^{-3}$</td>
<td>55</td>
</tr>
<tr>
<td>288</td>
<td>G</td>
<td>G (1.00)</td>
<td>G (0.99)</td>
<td>C (0.00)</td>
<td>Gp2</td>
<td>G96G</td>
<td>$4.83 \times 10^{-3}$</td>
<td>54</td>
</tr>
<tr>
<td>289</td>
<td>A</td>
<td>A (0.90)</td>
<td>A (0.79)</td>
<td>G (0.15)</td>
<td>Gp2</td>
<td>M97V*</td>
<td>$6.52 \times 10^{-4}$</td>
<td>40</td>
</tr>
<tr>
<td>291</td>
<td>G</td>
<td>G (0.60)</td>
<td>G (0.98)</td>
<td>A (0.00)</td>
<td>Gp2</td>
<td>M97T*</td>
<td>$7.04 \times 10^{-5}$</td>
<td>34</td>
</tr>
<tr>
<td>402</td>
<td>C</td>
<td>C (1.00)</td>
<td>C (0.98)</td>
<td>G (0.01)</td>
<td>Gp2</td>
<td>A134A</td>
<td>$8.29 \times 10^{-4}$</td>
<td>41</td>
</tr>
<tr>
<td>421</td>
<td>G</td>
<td>G (1.00)</td>
<td>G (0.96)</td>
<td>A (0.04)</td>
<td>Gp2</td>
<td>G141S*</td>
<td>$1.98 \times 10^{-68}$</td>
<td>11</td>
</tr>
<tr>
<td>472</td>
<td>A</td>
<td>A (1.00)</td>
<td>A (0.91)</td>
<td>G (0.03)</td>
<td>Gp2</td>
<td>T158A*</td>
<td>$1.13 \times 10^{-60}$</td>
<td>13</td>
</tr>
<tr>
<td>607</td>
<td>C</td>
<td>C (1.00)</td>
<td>C (0.99)</td>
<td>A (0.01)</td>
<td>Gp2</td>
<td>L203I*</td>
<td>$4.4 \times 10^{-4}$</td>
<td>38</td>
</tr>
<tr>
<td>Site</td>
<td>NVSL</td>
<td>Inoc</td>
<td>Maj</td>
<td>Min</td>
<td>Location</td>
<td>Amino Acid</td>
<td>p-value</td>
<td>R</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>----------</td>
<td>------------</td>
<td>---------</td>
<td>----</td>
</tr>
<tr>
<td>640</td>
<td>C</td>
<td>C (1.00)</td>
<td>C (0.98)</td>
<td>T (0.01)</td>
<td>Gp2; Gp3</td>
<td>H214Y*/ A6V*</td>
<td>$3.38 \times 10^{-2}$</td>
<td>77</td>
</tr>
<tr>
<td>813</td>
<td>G</td>
<td>G (1.00)</td>
<td>G (0.98)</td>
<td>C (0.01)</td>
<td>Gp3</td>
<td>A64P*</td>
<td>$1.12 \times 10^{-3}$</td>
<td>43</td>
</tr>
<tr>
<td>815</td>
<td>T</td>
<td>T (1.00)</td>
<td>T (0.98)</td>
<td>C (0.01)</td>
<td>Gp3</td>
<td>A64A</td>
<td>$1.11 \times 10^{-3}$</td>
<td>42</td>
</tr>
<tr>
<td>819</td>
<td>G</td>
<td>G (1.00)</td>
<td>G (0.98)</td>
<td>C (0.01)</td>
<td>Gp3</td>
<td>V66L*</td>
<td>$5.98 \times 10^{-3}$</td>
<td>56</td>
</tr>
<tr>
<td>833</td>
<td>C</td>
<td>C (1.00)</td>
<td>C (0.98)</td>
<td>A (0.01)</td>
<td>Gp3; Gp5</td>
<td>L73V*</td>
<td>$1.28 \times 10^{-2}$</td>
<td>66</td>
</tr>
<tr>
<td>840</td>
<td>C</td>
<td>C (1.00)</td>
<td>C (0.98)</td>
<td>T (0.01)</td>
<td>Gp3</td>
<td>L152R</td>
<td>$9.98 \times 10^{-103}$</td>
<td>10</td>
</tr>
<tr>
<td>876</td>
<td>G</td>
<td>G (1.00)</td>
<td>G (0.91)</td>
<td>T (0.03)</td>
<td>Gp3</td>
<td>D85Y*</td>
<td>$3.75 \times 10^{-153}$</td>
<td>8</td>
</tr>
<tr>
<td>878</td>
<td>C</td>
<td>A (0.55)</td>
<td>A (0.86)</td>
<td>C (0.03)</td>
<td>Gp3</td>
<td>E85D*</td>
<td>$4 \times 10^{-3}$</td>
<td>53</td>
</tr>
<tr>
<td>905</td>
<td>T</td>
<td>T (1.00)</td>
<td>T (0.98)</td>
<td>G (0.01)</td>
<td>Gp3</td>
<td>V94V</td>
<td>$2.44 \times 10^{-2}$</td>
<td>74</td>
</tr>
<tr>
<td>909</td>
<td>C</td>
<td>C (0.55)</td>
<td>T (0.66)</td>
<td>C (0.32)</td>
<td>Gp3</td>
<td>P96P</td>
<td>$2.13 \times 10^{-41}$</td>
<td>15</td>
</tr>
<tr>
<td>913</td>
<td>G</td>
<td>G (1.00)</td>
<td>G (0.99)</td>
<td>T (0.01)</td>
<td>Gp3</td>
<td>G97V*</td>
<td>$2.48 \times 10^{-2}$</td>
<td>75</td>
</tr>
<tr>
<td>1034</td>
<td>T</td>
<td>T (1.00)</td>
<td>T (0.99)</td>
<td>A (0.01)</td>
<td>Gp3</td>
<td>V137V</td>
<td>$5.42 \times 10^{-11}$</td>
<td>20</td>
</tr>
<tr>
<td>1178</td>
<td>C</td>
<td>C (1.00)</td>
<td>C (0.98)</td>
<td>G (0.01)</td>
<td>Gp3; Gp4</td>
<td>R135R; P4A*</td>
<td>$1.18 \times 10^{-4}$</td>
<td>35</td>
</tr>
<tr>
<td>1193</td>
<td>T</td>
<td>T (1.00)</td>
<td>T (0.99)</td>
<td>C (0.00)</td>
<td>Gp3; Gp4</td>
<td>S190S; L9L</td>
<td>$7.11 \times 10^{-3}$</td>
<td>59</td>
</tr>
<tr>
<td>1240</td>
<td>G</td>
<td>G (1.00)</td>
<td>G (0.98)</td>
<td>C (0.01)</td>
<td>Gp3; Gp4</td>
<td>S206T*; K24N*</td>
<td>$6.95 \times 10^{-3}$</td>
<td>58</td>
</tr>
<tr>
<td>1282</td>
<td>T</td>
<td>T (1.00)</td>
<td>T (0.89)</td>
<td>C (0.04)</td>
<td>Gp3; Gp4</td>
<td>L220P*; T38T</td>
<td>$3.22 \times 10^{-154}$</td>
<td>7</td>
</tr>
<tr>
<td>1442</td>
<td>C</td>
<td>C (1.00)</td>
<td>C (0.98)</td>
<td>T (0.01)</td>
<td>Gp4</td>
<td>H92Y*</td>
<td>$2.75 \times 10^{-3}$</td>
<td>60</td>
</tr>
<tr>
<td>1453</td>
<td>T</td>
<td>T (1.00)</td>
<td>T (0.99)</td>
<td>A (0.01)</td>
<td>Gp4</td>
<td>D95D</td>
<td>$3.07 \times 10^{-2}$</td>
<td>70</td>
</tr>
<tr>
<td>1490</td>
<td>G</td>
<td>G (1.00)</td>
<td>G (0.98)</td>
<td>T (0.01)</td>
<td>Gp4</td>
<td>E108**</td>
<td>$4.21 \times 10^{-6}$</td>
<td>31</td>
</tr>
<tr>
<td>1491</td>
<td>A</td>
<td>A (1.00)</td>
<td>A (0.98)</td>
<td>T (0.01)</td>
<td>Gp4</td>
<td>E105V*</td>
<td>$1.17 \times 10^{-2}$</td>
<td>65</td>
</tr>
<tr>
<td>1499</td>
<td>G</td>
<td>G (1.00)</td>
<td>G (0.99)</td>
<td>T (0.01)</td>
<td>Gp4</td>
<td>E111**</td>
<td>$4.08 \times 10^{-2}$</td>
<td>64</td>
</tr>
<tr>
<td>1501</td>
<td>A</td>
<td>A (1.00)</td>
<td>A (0.98)</td>
<td>G (0.01)</td>
<td>Gp4</td>
<td>E111E</td>
<td>$1.74 \times 10^{-3}$</td>
<td>46</td>
</tr>
<tr>
<td>1510</td>
<td>C</td>
<td>C (1.00)</td>
<td>C (0.94)</td>
<td>T (0.05)</td>
<td>Gp4</td>
<td>F114F</td>
<td>$9.54 \times 10^{-33}$</td>
<td>17</td>
</tr>
<tr>
<td>1517</td>
<td>A</td>
<td>A (1.00)</td>
<td>A (0.99)</td>
<td>G (0.01)</td>
<td>Gp4</td>
<td>I117V*</td>
<td>$3.88 \times 10^{-2}$</td>
<td>81</td>
</tr>
<tr>
<td>1523</td>
<td>G</td>
<td>G (1.00)</td>
<td>G (0.94)</td>
<td>A (0.07)</td>
<td>Gp4</td>
<td>G119S*</td>
<td>$4.05 \times 10^{-230}$</td>
<td>4</td>
</tr>
<tr>
<td>1538</td>
<td>A</td>
<td>A (0.60)</td>
<td>A (0.94)</td>
<td>G (0.02)</td>
<td>Gp4</td>
<td>I124V*</td>
<td>$0$</td>
<td>1</td>
</tr>
<tr>
<td>1553</td>
<td>G</td>
<td>G (0.90)</td>
<td>G (0.54)</td>
<td>A (0.44)</td>
<td>Gp4</td>
<td>I129I</td>
<td>$3.83 \times 10^{-203}$</td>
<td>5</td>
</tr>
<tr>
<td>1688</td>
<td>A</td>
<td>A (1.00)</td>
<td>A (0.95)</td>
<td>G (0.06)</td>
<td>Gp4</td>
<td>E174V*</td>
<td>$1.88 \times 10^{-302}$</td>
<td>3</td>
</tr>
<tr>
<td>1771</td>
<td>G</td>
<td>G (1.00)</td>
<td>G (0.99)</td>
<td>T (0.01)</td>
<td>Gp5; Gp5a</td>
<td>C19F*; V22V</td>
<td>$2.27 \times 10^{-4}$</td>
<td>36</td>
</tr>
<tr>
<td>1838</td>
<td>G</td>
<td>G (1.00)</td>
<td>G (0.99)</td>
<td>T (0.01)</td>
<td>Gp5; Gp5a</td>
<td>L41F*; D45Y*</td>
<td>$2.92 \times 10^{-2}$</td>
<td>76</td>
</tr>
<tr>
<td>1854</td>
<td>C</td>
<td>C (1.00)</td>
<td>C (0.98)</td>
<td>T (0.01)</td>
<td>Gp5; Gp5a</td>
<td>L47L; A50V*</td>
<td>$5.82 \times 10^{-4}$</td>
<td>39</td>
</tr>
<tr>
<td>1958</td>
<td>A</td>
<td>A (1.00)</td>
<td>A (0.92)</td>
<td>T (0.06)</td>
<td>Gp5</td>
<td>A81A</td>
<td>$7.39 \times 10^{-49}$</td>
<td>14</td>
</tr>
<tr>
<td>1964</td>
<td>C</td>
<td>C (1.00)</td>
<td>C (0.99)</td>
<td>T (0.01)</td>
<td>Gp5</td>
<td>T83T</td>
<td>$7.4 \times 10^{-6}$</td>
<td>32</td>
</tr>
<tr>
<td>2019</td>
<td>T</td>
<td>T (1.00)</td>
<td>T (0.91)</td>
<td>C (0.04)</td>
<td>Gp5</td>
<td>Y102H*</td>
<td>$1.18 \times 10^{-137}$</td>
<td>9</td>
</tr>
<tr>
<td>2094</td>
<td>C</td>
<td>C (1.00)</td>
<td>C (0.99)</td>
<td>T (0.01)</td>
<td>Gp5</td>
<td>L127F*</td>
<td>$3.7 \times 10^{-3}$</td>
<td>52</td>
</tr>
<tr>
<td>2126</td>
<td>T</td>
<td>T (1.00)</td>
<td>T (0.97)</td>
<td>A (0.02)</td>
<td>Gp5</td>
<td>S137S</td>
<td>$1.16 \times 10^{-65}$</td>
<td>12</td>
</tr>
<tr>
<td>2170</td>
<td>T</td>
<td>T (1.00)</td>
<td>T (0.98)</td>
<td>G (0.01)</td>
<td>Gp5</td>
<td>L152R*</td>
<td>$2.36 \times 10^{-3}$</td>
<td>49</td>
</tr>
<tr>
<td>2177</td>
<td>C</td>
<td>C (1.00)</td>
<td>C (0.96)</td>
<td>T (0.01)</td>
<td>Gp5</td>
<td>R154R</td>
<td>$1.06 \times 10^{-165}$</td>
<td>6</td>
</tr>
<tr>
<td>2190</td>
<td>G</td>
<td>G (1.00)</td>
<td>G (0.98)</td>
<td>A (0.01)</td>
<td>Gp5</td>
<td>V159I*</td>
<td>$2.2 \times 10^{-38}$</td>
<td>16</td>
</tr>
<tr>
<td>2200</td>
<td>A</td>
<td>A (1.00)</td>
<td>A (0.99)</td>
<td>T (0.01)</td>
<td>Gp5</td>
<td>E162V*</td>
<td>$4.87 \times 10^{-2}$</td>
<td>83</td>
</tr>
<tr>
<td>2205</td>
<td>G</td>
<td>G (1.00)</td>
<td>G (0.99)</td>
<td>A (0.01)</td>
<td>Gp5</td>
<td>G164R*</td>
<td>$3.73 \times 10^{-2}$</td>
<td>80</td>
</tr>
</tbody>
</table>
Table 2.5 (continued)

<table>
<thead>
<tr>
<th>Site</th>
<th>NVSL</th>
<th>Inoc</th>
<th>Maj</th>
<th>Min</th>
<th>Location</th>
<th>Amino Acid</th>
<th>p-value</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>2218</td>
<td>A</td>
<td>A</td>
<td>1.00</td>
<td>0.99</td>
<td>G</td>
<td>Gp5</td>
<td>E168G*</td>
<td>4.53 x 10^{-2}</td>
</tr>
<tr>
<td>2229</td>
<td>C</td>
<td>C</td>
<td>1.00</td>
<td>0.98</td>
<td>G</td>
<td>Gp5</td>
<td>H172D*</td>
<td>1.66 x 10^{-2}</td>
</tr>
<tr>
<td>2314</td>
<td>T</td>
<td>T</td>
<td>1.00</td>
<td>0.99</td>
<td>G</td>
<td>Gp5; M</td>
<td>L200R*; S4S</td>
<td>1.84 x 10^{-3}</td>
</tr>
<tr>
<td>2349</td>
<td>A</td>
<td>A</td>
<td>1.00</td>
<td>0.96</td>
<td>T</td>
<td>M</td>
<td>Q16L*</td>
<td>7.21 x 10^{-9}</td>
</tr>
<tr>
<td>2404</td>
<td>A</td>
<td>A</td>
<td>1.00</td>
<td>0.96</td>
<td>G</td>
<td>M</td>
<td>L34L</td>
<td>1.81 x 10^{-28}</td>
</tr>
<tr>
<td>2427</td>
<td>T</td>
<td>T</td>
<td>1.00</td>
<td>0.99</td>
<td>C</td>
<td>M</td>
<td>L42S*</td>
<td>2.55 x 10^{-3}</td>
</tr>
<tr>
<td>2459</td>
<td>T</td>
<td>T</td>
<td>1.00</td>
<td>0.99</td>
<td>C</td>
<td>M</td>
<td>C53R*</td>
<td>2.6 x 10^{-4}</td>
</tr>
<tr>
<td>2460</td>
<td>G</td>
<td>G</td>
<td>1.00</td>
<td>0.98</td>
<td>T</td>
<td>M</td>
<td>C53F*</td>
<td>2.19 x 10^{-2}</td>
</tr>
<tr>
<td>2497</td>
<td>T</td>
<td>T</td>
<td>1.00</td>
<td>0.98</td>
<td>C</td>
<td>M</td>
<td>F65F</td>
<td>2.31 x 10^{-2}</td>
</tr>
<tr>
<td>2518</td>
<td>G</td>
<td>G</td>
<td>1.00</td>
<td>0.98</td>
<td>A</td>
<td>M</td>
<td>A72A</td>
<td>1.14 x 10^{-6}</td>
</tr>
<tr>
<td>2778</td>
<td>G</td>
<td>G</td>
<td>1.00</td>
<td>0.98</td>
<td>T</td>
<td>M</td>
<td>R159I*</td>
<td>0</td>
</tr>
<tr>
<td>2802</td>
<td>T</td>
<td>T</td>
<td>1.00</td>
<td>0.97</td>
<td>G</td>
<td>M</td>
<td>V167G*</td>
<td>9.68 x 10^{-3}</td>
</tr>
<tr>
<td>2803</td>
<td>A</td>
<td>A</td>
<td>1.00</td>
<td>0.98</td>
<td>G</td>
<td>M</td>
<td>V167V</td>
<td>1.31 x 10^{-2}</td>
</tr>
<tr>
<td>2820</td>
<td>C</td>
<td>C</td>
<td>1.00</td>
<td>0.97</td>
<td>A</td>
<td>M</td>
<td>A173D*</td>
<td>7.83 x 10^{-3}</td>
</tr>
</tbody>
</table>

2.3.4 Hellams test

In the methods section, we conjectured that \( \Pr(X_t \geq x_0) \leq \max\{x_0, 1 - x_0\} \), where \( x_0 \) and \( X_t \) are the relative allele frequencies at an earlier and later time point, respectively. We then used the conjecture to propose the Hellams test for selection. In this section, we provide evidence to support the conjecture, demonstrate the utility of the test in a small simulation, and apply the test to the host dataset.

While we do not have a formal proof of the conjecture, we empirically confirmed its veracity for all the 6,370 simulation conditions in chapter 3, where \( x_0 > 0.5 \). There were no cases where \( \Pr(X_t \geq x_0) \) exceeded \( x_0 \).

We then undertook a small simulation study to understand the power of the Hellams test when sites were evolving by the Wright-Fisher model with or without selection and \( N_e = 200 \) and \( t = 50 \) generations. We varied the selection coefficient \( s \in \{0, 0.01, 0.1\} \), the initial allele frequency \( x_0 \in \{0.5, 0.7, 0.9\} \), the sequencing coverage in \{10, 20, 50\}, and the number of replicates \( n \in \{5, 20, 100\} \). Actual NGS coverage in the PRRSV RNA-seq data often exceeds the upper limit 50 used in simulation by orders of magnitude, but these are typical coverage levels in DNA-seq
Table 2.6: Simulation results for the Hellams test. The power to reject $H_0 : s \leq 0$ at significance level $\alpha = 0.05$ for various choices of selection coefficient $s$, initial allele frequency $x_0$, sequencing coverage, and number of replicates, which may be experimental units or time intervals, or a combination of both. When $s = 0$, there is no selection.

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Coverage 10</th>
<th>Coverage 20</th>
<th>Coverage 50</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n = 5$</td>
<td>$n = 20$</td>
<td>$n = 100$</td>
</tr>
<tr>
<td>$s = 0$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_0 = 0.5$</td>
<td>0.000</td>
<td>0.020</td>
<td>0.000</td>
</tr>
<tr>
<td>$x_0 = 0.7$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>$x_0 = 0.9$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>$s = 0.01$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_0 = 0.5$</td>
<td>0.000</td>
<td>0.080</td>
<td>0.440</td>
</tr>
<tr>
<td>$x_0 = 0.7$</td>
<td>0.000</td>
<td>0.060</td>
<td>0.140</td>
</tr>
<tr>
<td>$x_0 = 0.9$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>$s = 0.1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_0 = 0.5$</td>
<td>0.020</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$x_0 = 0.7$</td>
<td>0.020</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$x_0 = 0.9$</td>
<td>0.000</td>
<td>0.020</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Experiments and allow us to explore the effect of lower coverage. The number of replicates is a combination of the number of experimental units, if the selection coefficient is assumed to be constant across units, and the number of time intervals. In our real dataset, there are 13 pigs each with 3 serial time intervals, leading to $n = 39$. In the total PHGC study, there are hundreds of infected pigs. Table 2.6 reports the power computed for 50 replicates per simulation condition. Appropriately, there is no power to detect selection when there is no selection. When selection is moderate, $s = 0.01$, then there is good power (above 0.80) when $x_0 = 0.5$, coverage is high, and the number of replicates is high. When selection is strong, $s = 0.1$, there is good coverage except when there are only $n = 5$ replicates or even $n = 20$ replicates, but also $x_0 = 0.9$. It is clear that there is little power to detect selection for alleles that are already nearing fixation. In addition, longer time series, or more experimental units can greatly increase the power.

Finally, we applied the proposed test to the host data, but found no evidence of selection. Table 2.7 merely reports $p$-values less than 0.5, and there has been no correction for multiple testing. The test has little power under the conditions of these data, particularly due to the fact that the major allele relative frequency almost always exceeds 0.9, and usually exceeds 0.98 (Table 2.5).
Table 2.7: Hellams test applied RNA-seq data of host study, showing only results where \( p \)-value was below 0.5. No sites were significantly selected.

<table>
<thead>
<tr>
<th>Pig(^a)</th>
<th>Site</th>
<th>( H_0^b )</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>83</td>
<td>( s &gt; 0 )</td>
<td>0.49</td>
</tr>
<tr>
<td>3</td>
<td>289</td>
<td>( s &gt; 0 )</td>
<td>0.29</td>
</tr>
<tr>
<td>5</td>
<td>289</td>
<td>( s &lt; 0 )</td>
<td>0.17</td>
</tr>
<tr>
<td>7</td>
<td>813</td>
<td>( s &gt; 0 )</td>
<td>0.48</td>
</tr>
<tr>
<td>12</td>
<td>878</td>
<td>( s &lt; 0 )</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>909</td>
<td>( s &lt; 0 )</td>
<td>0.24</td>
</tr>
<tr>
<td>5</td>
<td>909</td>
<td>( s &lt; 0 )</td>
<td>0.29</td>
</tr>
<tr>
<td>6</td>
<td>909</td>
<td>( s &lt; 0 )</td>
<td>0.20</td>
</tr>
<tr>
<td>10</td>
<td>1538</td>
<td>( s &gt; 0 )</td>
<td>0.37</td>
</tr>
<tr>
<td>5</td>
<td>1553</td>
<td>( s &gt; 0 )</td>
<td>0.47</td>
</tr>
<tr>
<td>6</td>
<td>1553</td>
<td>( s &gt; 0 )</td>
<td>0.40</td>
</tr>
<tr>
<td>10</td>
<td>1553</td>
<td>( s &gt; 0 )</td>
<td>0.29</td>
</tr>
<tr>
<td>12</td>
<td>1553</td>
<td>( s &gt; 0 )</td>
<td>0.28</td>
</tr>
<tr>
<td>13</td>
<td>1553</td>
<td>( s &gt; 0 )</td>
<td>0.30</td>
</tr>
<tr>
<td>all</td>
<td>1553</td>
<td>( s &gt; 0 )</td>
<td>0.17</td>
</tr>
<tr>
<td>5</td>
<td>1958</td>
<td>( s &lt; 0 )</td>
<td>0.44</td>
</tr>
</tbody>
</table>

\(^a\)Pig number (1–13) or “all” if the test combined all pig data.

\(^b\)A positive selection coefficient on the major allele frequency, \( s > 0 \) implies the major allele should tend to increase in frequency, otherwise decrease.

In addition, as noted in the methods, the proposed test has no ability to detect selection against major alleles or selection for minor alleles.

### 2.3.5 Linkage disequilibrium

The sites and therefore tests implemented in the preceding sections are not independent. Within genes, the long Sanger reads provide information about association of mutations along the sequence. We perform a Fisher exact test for linkage equilibrium between all pairs of selected sites in all pig/dpi combinations. At the unadjusted significance level 0.05, there are two clusters of sites in ORF2–6 that reject the null hypothesis of linkage equilibrium for at least one pig/dpi combination: (1) sites 1553 and 1795 (\( p \)-value \( 2.86 \times 10^{-2} \)) and (2) sites 711, 1050, and 1838 (\( p \)-values between pairs \( 711, 1050 \), \( 711, 1838 \), and \( 1050, 1838 \) are \( 6.82 \times 10^{-3} \), \( 1.04 \times 10^{-4} \), and \( 2.23 \times 10^{-2} \)).
Interestingly, all these sites, but 1795, were confirmed by the host study. It is likely that only one of the pair 1538 and 1553 is selected, perhaps the nonsynonymous change at 1538. For the other cluster, it is unclear whether 291, 909, or both might be selected. The long range association with 1795 may be an artefact of the special nature of the 1795, noted earlier, or it could reflect long-range physical interactions, apparently in the RNA, since the change at 1795 is nonsynonymous. In the \textit{nsp2} gene, sites 1167, 1223 and 1645 are linked with \textit{p}-values $4.35 \times 10^{-3}$, $9.61 \times 10^{-6}$ and $4.29 \times 10^{-1}$; sites 1167 and 1645 are indirectly linked via their connection to 1223.

2.4 Discussion

Most existing tests of selection in temporal data require an estimate or simultaneous estimation of the effective population size \( N_e \), since this parameter has a large effect on the rate of change in relative allele frequencies. Unfortunately, genetic data contain very weak information about \( N_e \), leading to imprecise estimates and leaving little power to detect selection. The Hellams test, which requires no estimate of \( N_e \), is not immune to this conundrum. It is not a particularly powerful test, especially under the data conditions most prevalent in the PRRSV datasets, in particular, the high initial relative allele frequencies.

Nevertheless, there is some evidence gathered from comparing the total 18 pigs in this study. Site 30 is detected in both datasets, and there does appear to be a notable shift in all pigs toward the G allele at this site, although it is already the major allele in the inoculum. There is a similar shift toward an already major allele at many of the confirmed sites (all together, sites 30, 159, 291, 1050, 1538). The Hellams test, however, is not impressed by the number of replicated shifts, but is underpowered for these sites where the initial relative allele frequency starts high. Another set of sites shift from one allele to another, but not necessarily consistently across pigs (sites 909, 1523, 1553, 1838). These sites are polymorphic to start and genetic drift is a plausible explanation for the movement. Hellams test has reasonable power for some of these sites, and indeed agrees that the changes are not sufficiently consistent across pigs or time points to warrant rejection of the null hypothesis of neutral evolution.
The cases most likely to be selected (sites 30, 159, 291, 1050, and 1538), if selected, are an example of purifying selection for a particular preferred allele. Selection that acts identically across all pigs is expected to be of the purifying type: there is simply one allele that works better *in vivo*. Virus grown in cell culture are known to adapt, accumulating changes relative to field isolates (de Abin et al., 2008; Schommer, 2000; Chang et al., 2017). Since the inoculum was grown in culture, we therefore expect some sites to respond to purifying selection once introduced into the pig. In contrast, selection for genetic diversification, in order to avoid an immune response, for example, would be pig specific. CMH has low power to detect such selection, but cannot confirm whether it is selection. The Hellams test can be useful, but it has insufficient power in this study given the limited length of the time series in each pig, at most three time intervals. Furthermore, there should be little immune activity in the pigs of the host study since the sampled time period represents only the first two weeks of an infection, before an adaptive immune response has been mounted (Islam et al., 2017). The last time point in the virus study could provide evidence of immune-mediated selection, but one time interval provides insufficient power for all tests.
CHAPTER 3. ENHANCED FITR TEST FOR SELECTION

3.1 Introduction

As sequence data become more and more ubiquitous, there is growing interest in detecting natural selection from temporal genetic samples. The selection pressures acting on pathogens during key transitions, for example, may reveal the events that accompany rare cross-species transmission events (Poss et al., 2006), between host transmission events (Carlson et al., 2014), and evasion of within-host immune pressures (Gounder et al., 2015). Detailed knowledge of selection events that lead to effective immune responses during natural infections may suggest better vaccines or vaccination strategies (Bhiman et al., 2015). Most temporal data arise in the context of rapidly evolving species, but selection that acts on slower time scales can also generate temporal samples. For example, historical data available in ancient DNA samples may one day reveal the selection events that gave rise to modern species (Malaspinas, 2016).

Most tests or estimates of selection require simultaneous estimation of nuisance parameters, like the effective population size \( N_e \) or the elapsed evolutionary time \( \Delta \tau \) between sample points. Nishino (2013) recently introduced a test, the FITR test, that capitalizes on the breadth of data available at each time point to avoid explicit estimation of these nuisance parameters. Genome-wide data obtained from population samples provide abundance information for a multitude of loci. If some of these loci are occupied by neutral alleles that evolve independently of each other, then they can provide indirect estimates of \( N_e \) and \( \Delta \tau \). The test relies on the Gaussian approximation to the relative frequency increment distribution, where the relative frequency increment \( \Delta x_i = x_i - x_{i-1} \) is the change in allele relative frequency across an interval of time.

Models for the evolution of allele relative frequencies in time are notoriously intractable. The Gaussian approximation is not valid for long time lapses, small effective population sizes, or allele frequencies near the boundaries \( \{0, 1\} \) of their domain. New and better approximations are actively
sought (Tataru et al., 2016). Here, we introduce an approach that extends the applicability of the FITR test by transforming the scaled frequency increments toward normality. We present the model with our proposed transformations and use simulation to show the Box-Cox transformation performs best. The transformation parameters depend on the initial allele frequency \( x_i \), so we next propose a method to combine information from neutral loci with varying initial allele frequencies. The final proposed test with transformation is shown to have superior performance compared to the FITR test.

### 3.2 Model

Suppose we track the population major allele relative frequency \( x_{ri} \) at locus \( r \) at times \( \tau_i, i \in \{0, 1, \ldots, L\} \). Under the Gaussian approximation to the neutral Wright-Fisher model, the change in allele frequency \( \Delta x_{ri} = x_{ri} - x_{r,i-1} \), conditional on the initial allele frequency \( x_{r,i-1} \), follows a normal distribution. Specifically, the scaled value

\[
y_{ri} := \frac{\Delta x_{ri}}{\sqrt{x_{r,i-1}(1 - x_{r,i-1})}} \sim \mathcal{N}\left(0, \frac{\delta \Delta \tau_i}{2N_i}\right),
\]

where \( N_i \) is the variance effective population size, \( \Delta \tau_i \) is the elapsed time over interval \( (\tau_{i-1}, \tau_i) \), and \( \delta \) generations per unit time is a scaling factor. If \( N_i = N \) is constant in time, the Frequency Increment Test statistic (Feder et al., 2014)

\[
t_{\text{FIT}} = \frac{y_r}{s_r / \sqrt{L}}
\]

with sample statistics

\[
y_r = \frac{1}{L} \sum_{i=1}^{L} y_{ri} \quad \text{and} \quad s_r^2 = \frac{1}{L-1} \sum_{i=1}^{L} (y_{ri} - y_r)^2,
\]

approximately follows a \( t \) distribution with \( L - 1 \) degrees of freedom.

When populations vary in size, the effective population size \( N_i \) varies across time increment \( i \), so it can no longer be estimated via \( s_r^2 \). Nishino (2013) proposes to instead measure \( R \) independent reference loci known to be evolving neutrally. Then, if \( y_{ri} \) are the scaled increments for the \( r \)th
reference locus and 0 indexes the test locus, the FIT with reference loci test statistic for each time
interval,

\[ t_{\text{FITR},i} = \frac{y_{0i}}{\sqrt{\frac{1}{R} \sum_{r=1}^{R} y_{ri}^2}}, \]

are approximately \( t_R \)-distributed, and

\[ t_{\text{FITR}} = \sum_{i=1}^{L} t_{\text{FITR},i} \]

has the distribution of the sum of \( L \) independent \( t \) statistics with \( R \) degrees of freedom. Since the
\( t_{\text{FITR},i} \) are iid and do not depend on the time interval \( [\tau_{i-1}, \tau_i] \), we now drop the dependence on
\( i \). However, we still track multiple scaled allele increments \( y_1, y_2, \ldots, y_R \) with corresponding initial
allele frequencies \( x_1, x_2, \ldots, x_R \) from \( R \) distinct reference loci.

Both tests rely on the time- and space-continuous Gaussian approximation to the underlying
discrete genetic model. In particular, the Gaussian approximation is not valid as \( x_r \) approaches
the boundaries, \( \{0, 1\} \), especially as the lapsed time \( \Delta \tau \) increases and effective population size
\( N \) decreases. When \( N \) remains sufficiently large, however, it is still reasonable to approximate
the distribution of the scaled increments \( y_r \) with a continuous, perhaps non-normal, approxima-
tion. Indeed, others have successfully used truncated Normal distributions, Beta distributions and
mixtures of Beta distributions and point masses (Tataru et al., 2016) to approximate the unscaled
frequency \( x_{ri} \) distribution conditional on \( x_{ri-1} \) and \( \Delta \tau_i \). We propose here to use normalizing trans-
formations to extend the usefulness of the FITR tests to scenarios where a continuous, non-normal
distributions well-approximate the distribution of \( y_r \).

3.2.1 Transformations

Not only does \( y_r \) exist on a finite interval, there is positive probability that the boundary is
actually achieved. Therefore,

\[ y_r \in \left[ \frac{-x_r}{\sqrt{x_r(1-x_r)}}, \frac{1-x_r}{\sqrt{x_r(1-x_r)}} \right] := [a_r, b_r], \]

lies in a closed interval, and the right boundary is substantially more likely to be reached for \( x_r \)
already close to 1, small \( N \), and long time lapses \( \Delta \tau \). To model positive mass on the boundaries,
we treat the distribution of $y_r$ as continuous on $(a_r, b_r)$ with point masses, $\gamma_{ra}$ at $a_r$ and $\gamma_{rb}$ at $b_r$. Let $u_r = I(y_r \neq a_r)$ and $v_r = I(y_r \neq b_r)$ be observable binary variables indicating observations on the boundaries. Observation $y_r$ is generated by first sampling $(u_r, v_r, 1 - u_r - v_r) \sim \text{Multinoulli}(\gamma_{ra}, \gamma_{rb}, 1 - \gamma_{ra} - \gamma_{rb})$. Conditional on $u_r = v_r = 1$, $y_r$ is drawn from some continuous distribution that is defined by the assumption that there exists a transformation $g(\cdot; \lambda)$, a function of parameter vector $\lambda$, that can transform $y_r$ such that, conditionally,

$$w_r := g(y_r; \lambda_r) \mid y_r \in (a_r, b_r) \sim N(\mu_r, \sigma_r^2).$$

Because this process is dependent on initial allele relative frequency $x_r$, we must account for the possibility that the point masses, the transformation parameters, as well as the mean and variance of the resulting normal distribution may depend on the locus $r$.

Unconditionally, the observed data $y_r$ have density

$$f_r(y_r | \lambda_r, \mu_r, \sigma_r^2, \gamma_{ra}, \gamma_{rb}) = \begin{cases} (1 - \gamma_{ra} - \gamma_{rb})\phi \left( \frac{g(y_r; \lambda_r) - \mu_r}{\sigma_r} \right) \left| \frac{dg(y_r; \lambda_r)}{dy} \right|_{y=y_r}, & y_r \in (a_r, b_r) \\ \gamma_{ra}, & y_r = a_r \\ \gamma_{rb}, & y_r = b_r \\ 0, & \text{otherwise,} \end{cases} \quad (3.1)$$

where $\phi$ is the standard normal pdf. The log-likelihood function given the observed data $y = (y_1, y_2, \ldots, y_R)$ is

$$l(\theta \mid y) = \sum_{r=1}^R l_r(\theta_r \mid y_r) = \sum_{r=1}^R \ln f_r(y_r \mid \lambda_r, \mu_r, \sigma_r^2, \gamma_{ra}, \gamma_{rb})$$

$$= \sum_{r=1}^R u_r v_r \left\{-\frac{1}{2} \ln(2\pi\sigma_r^2) - \frac{[g(y_r; \lambda_r) - \mu_r]^2}{2\sigma_r^2} + \ln \left| \frac{dg(y_r; \lambda_r)}{dy} \right|_{y=y_r} \right\} \quad (3.2)$$

$$+ \sum_{r=1}^R [u_r v_r \ln(1 - \gamma_{ra} - \gamma_{rb}) + (1 - u_r) \ln \gamma_{ra} + (1 - v_r) \ln \gamma_{rb}],$$

where the complete vector of parameters is $\theta = (\lambda, \mu, \sigma^2, \gamma_a, \gamma_b)$ and $\theta_r = (\lambda_r, \mu_r, \sigma_r^2, \gamma_{ra}, \gamma_{rb})$ are the parameters for the $r$th reference locus. Note, the likelihood can be written as two parts that can be independently maximized: $l_{\text{out}}(\gamma_a, \gamma_b \mid u, v)$ involving $\gamma_a$ and $\gamma_b$ and $l_{\text{in}}(\lambda, \mu, \sigma^2 \mid y)$ involving $\lambda$ from the transformation and $\mu$ and $\sigma^2$ from the normal distribution.
Table 3.1: Transformations used to normalize $t = \lambda_r y_r$, with $\lambda_r = b_r + 1 \times 10^{-6}$. Since $\lambda_r y_r$ is strictly positive, we only use two conditions of the Yeo-Johnson transformation (Yeo and Johnson, 2000). In practice, both $t$ and $\lambda_1$ depend on locus $r$, but are written without indices below.

<table>
<thead>
<tr>
<th>Transformation</th>
<th>$g(t; \lambda_1)$</th>
<th>$\frac{\partial g(t; \lambda_1)}{\partial t}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Box-Cox (BC)</td>
<td>$\frac{t^\lambda_1 - 1}{\lambda_1}$, $\lambda_1 \neq 0$</td>
<td>$t^{\lambda_1 - 1}$, $\lambda_1 \neq 0$</td>
</tr>
<tr>
<td>(Box and Cox, 1964)</td>
<td>log $t$, $\lambda_1 = 0$</td>
<td>$\frac{1}{t}$, $\lambda_1 = 0$</td>
</tr>
<tr>
<td>Yeo-Johnson (YJ)</td>
<td>$\frac{(t+1)^{\lambda_1} - 1}{\lambda_1}$, $\lambda_1 \neq 0$</td>
<td>$(t + 1)^{\lambda_1 - 1}$, $\lambda_1 \neq 0$</td>
</tr>
<tr>
<td>(Yeo and Johnson, 2000)</td>
<td>log $(t + 1)$, $\lambda_1 = 0$</td>
<td>$\frac{1}{t + 1}$, $\lambda_1 = 0$</td>
</tr>
<tr>
<td>Manly (MAN)</td>
<td>$\frac{e^{\lambda_1 t}}{\lambda_1}$, $\lambda_1 \neq 0$</td>
<td>$e^{\lambda_1 t}$, $\lambda_1 \neq 0$</td>
</tr>
<tr>
<td>(Manly, 1976)</td>
<td>$t$ $\lambda_1 = 0$</td>
<td>$1$, $\lambda_1 = 0$</td>
</tr>
<tr>
<td>Inverse Hyperbolic Sine (IHS)</td>
<td>$\log \left( \frac{\lambda_1 t + \sqrt{\lambda_1^2 t^2 + 1}}{\lambda_1} \right)$, $\lambda_1 \neq 0$</td>
<td>$\frac{1}{\sqrt{\lambda_1^2 t^2 + 1}}$, $\lambda_1 \neq 0$</td>
</tr>
<tr>
<td>(Johnson, 1949)</td>
<td>$t$, $\lambda_1 = 0$</td>
<td>$1$, $\lambda_1 = 0$</td>
</tr>
</tbody>
</table>

The transformations we consider are two-parameter transformations where $\lambda_r$ is a fixed shift parameter. We choose $\lambda_r = b_r + \delta$ for small positive $\delta$ ($1 \times 10^{-6}$ in practice) such that $\lambda_r y_r$ is strictly positive. This choice allows use of the Box-Cox transformation (Box and Cox, 1964), but also aligns the data such that the intent of the transformations to correct positive skew works for the shifted and negatively skewed scaled changes in major allele frequency (see Fig. 3.1). Estimating $\lambda_r$ only slightly improves normality overall, and for a few datasets, results in pathological optimization for Box-Cox because the Jacobian becomes unbounded as $\lambda_r$ approaches $\min\{y_r\}$ when $\lambda_1 < 1$ (data not shown). The complete list of transformations we consider and their derivatives are listed below in Table 3.1. When using the Box-Cox transformation, $\lambda_r y_r | u_r = v_r = 1 \overset{iid}{\sim} PN(\lambda_1, \mu_r, \sigma_r^2)$ follow a power normal distribution (Freeman and Modarres, 2006).

### 3.2.2 Simulation

Data were simulated as independent realizations of the neutral Wright-Fisher Markov chain over $T$ generations initialized with allele frequency $x$ and assuming effective population size $N$ (Table 3.2). Simulation A, a full factorial design over $N \in \{200, 1000\}$, $T \in \{10, 20, \ldots, N\}$, and $x \in \{0.50, 0.51, \ldots, 0.99\}$, was used to investigate the normality of scaled increment $y$ pre- and post-transformation. Simulations A1 and A2 are highlighted subsets of A focusing on two
Figure 3.1: Box-Cox transformation demonstrated for 2,316 observations, including fixations, of the scaled change in allele frequency with effective population size \( N = 1,000 \), number of generations \( T = 70 \), and initial allele frequency \( x = 0.95 \). The original data \( y \) (x-axis, blue) have a point mass of \( \gamma_B = 0.24 \) at about \( y = 0.23 \) and are highly skewed left. The transformed data \( w \) (y-axis, red) retain the point mass, now at \( w = 3.42 \), but are near normal elsewhere.

combinations of \( (N, T) \in \{(200, 40), (1000, 200)\} \) with \( x \in \{0.55, 0.67, 0.73, 0.90\} \) and were used to illustrate detailed results.

In simulation A, a simulation condition is defined by a single choice of \( N, T, \) and \( x \). Our final method takes mixed samples of neutral loci with the same \( N \) and \( T \), but varying \( x_r \) for locus \( r \). To simulate such data, we take the two \( (N, T) \) settings from A1 and A2 and simulate loci with \( x_r \) sampled uniformly from the non-fixed initial allele frequencies possible in a population of size \( N \), namely \( \{1, 2, \ldots, N - 1\}/N \).

We also simulate true positive loci subject to selection, where the Markov chain transition probabilities for allele count \( C_t \) at generation \( t \) are given by

\[
C_t \sim \text{Bin} \left( N, \frac{(1 + s)C_{t-1}}{N + sC_{t-1}} \right)
\]

for a selection coefficient \(-1 < s\). In simulation B1, we simulate positive selection \( s \in \{0.001, 0.01, 0.05, 0.1\} \), where sample statistics in the right tail provides evidence of non-neutrality. In simulation B2, we simulate negative selection \( s \in \{-0.001, -0.01, -0.05, -0.1\} \), where sample statistics in the
left tail indicate non-neutrality. For these simulations, we sample \( n_s = 10,000 \) loci in order to accurately estimate true and false positive rates.

**Table 3.2: Simulation settings.** Size is the number of simulation conditions, Replicates is the number of replicates per simulation setting. Internal replicates generate \( y_r \in (a_r, b_r) \), excluding the boundaries.

<table>
<thead>
<tr>
<th>Sim.</th>
<th>( N )</th>
<th>( T )</th>
<th>( x )</th>
<th>Selection</th>
<th>Size</th>
<th>Replicates</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>{200, 1000}</td>
<td>{10, 20, \ldots, N}</td>
<td>{0.50, 0.51, \ldots, 0.99}</td>
<td>( s = 0 )</td>
<td>6,000</td>
<td>1,000 internal</td>
</tr>
<tr>
<td>A1</td>
<td>{200}</td>
<td>{40}</td>
<td>{0.55, 0.67, 0.73, 0.90}</td>
<td>( s = 0 )</td>
<td>8</td>
<td>1,000 internal</td>
</tr>
<tr>
<td>A2</td>
<td>{1000}</td>
<td>{200}</td>
<td>{0.55, 0.67, 0.73, 0.90}</td>
<td>( s = 0 )</td>
<td>8</td>
<td>1,000 internal</td>
</tr>
<tr>
<td>B1.1</td>
<td>A1</td>
<td>Unif{1, 2, \ldots, N - 1}/N</td>
<td>( s = 0 )</td>
<td>1</td>
<td>10,000 total</td>
<td></td>
</tr>
<tr>
<td>B1.2</td>
<td>A1</td>
<td>Unif{1, 2, \ldots, N - 1}/N</td>
<td>( s \in {0.001, 0.01, 0.05, 0.1} )</td>
<td>4</td>
<td>10,000 total</td>
<td></td>
</tr>
<tr>
<td>B1.3</td>
<td>A1</td>
<td>Unif{1, 2, \ldots, N - 1}/N</td>
<td>( s \in {-0.001, -0.01, -0.05, -0.1} )</td>
<td>4</td>
<td>10,000 total</td>
<td></td>
</tr>
<tr>
<td>B2.1</td>
<td>A2</td>
<td>Unif{1, 2, \ldots, N - 1}/N</td>
<td>( s = 0 )</td>
<td>1</td>
<td>10,000 total</td>
<td></td>
</tr>
<tr>
<td>B2.2</td>
<td>A2</td>
<td>Unif{1, 2, \ldots, N - 1}/N</td>
<td>( s \in {0.001, 0.01, 0.05, 0.1} )</td>
<td>4</td>
<td>10,000 total</td>
<td></td>
</tr>
<tr>
<td>B2.3</td>
<td>A2</td>
<td>Unif{1, 2, \ldots, N - 1}/N</td>
<td>( s \in {-0.001, -0.01, -0.05, -0.1} )</td>
<td>4</td>
<td>10,000 total</td>
<td></td>
</tr>
</tbody>
</table>

### 3.2.3 Normalization with constant initial allele frequency \( x_r \)

Simulation A produces \( n_s = 1,000 \) replicate loci under each fixed \((N, T, x)\) condition. When loci share the same initial allele frequencies, \( x_r = x \) for all \( r \), as well as \( N \) and \( T \), then \( \lambda_r = \lambda \), \( \mu_r = \mu \), \( \sigma_r = \sigma \), bounds \( a, b \), and point masses \( \gamma_ra = \gamma_a, \gamma_rb = \gamma_b \) no longer depend on reference locus \( r \). Let \( y^{(1)}, y^{(2)}, \ldots, y^{(n_s)} \) be the simulated scaled allele frequency changes. Then, the maximum likelihood estimates of the mass on the boundary are \( \hat{\gamma}_a = \frac{1}{n_s} \sum_{i=1}^{n_s} (1 - u^{(i)}) \) and \( \hat{\gamma}_b = \frac{1}{n_s} \sum_{i=1}^{n_s} (1 - v^{(i)}) \), where \( 1 - u^{(i)} \) and \( 1 - v^{(i)} \) indicate \( y^{(i)} \) is on the left or right boundary. Similarly, at the global maximum of Eq. (3.2), \( \mu(\lambda_1) = \frac{1}{n_s} \sum_{i=1}^{n_s} u^{(i)} v^{(i)} z^{(i)}(\lambda_1) \) and \( \sigma^2(\lambda_1) = \frac{1}{n_s} \sum_{i=1}^{n_s} u^{(i)} v^{(i)} [z^{(i)}(\lambda_1) - \mu(\lambda_1)]^2 \), with \( z^{(i)}(\lambda_1) = g(y^{(i)}; \lambda_1, \lambda_2) \) the transformed data, are both trivial functions of the unknown transformation parameter \( \lambda_1 \) (recall \( \lambda_2 \) is a fixed, known parameter). The MLE of \( \lambda_1 \) can be found using
one-dimensional optimization of the profile log likelihood
\[
\hat{\lambda}_1 = \arg\max_{\lambda_1} \sum_{i=1}^{n_s} u^{(i)}_i v^{(i)} \left[ \ln \phi \left( \frac{z^{(i)}(\lambda_1) - \mu(\lambda_1)}{\sigma(\lambda_1)} \right) + \ln \left| \frac{dg(y; \lambda_1)}{dy} \right|_{y=y^{(i)}} \right],
\]
where \(\phi(\cdot)\) is the standard normal density.

### 3.2.4 Normalization with variable initial allele frequency \(x_r\)

While all reference loci observed over one time lapse share the same effective population size \(N\) and time lapse \(\Delta \tau\), it is not generally possible to sample \(R\) reference loci with the same, or maybe even similar, initial allele frequencies. To transform scaled allele frequency estimates with varying initial allele frequencies, we need to account for the unknown dependence of \(\lambda_{r1}, \mu_r, \sigma^2_r, \gamma_{ra},\) and \(\gamma_{rb}\) on \(x_r\). Because we know of no theoretical relationship, we choose to model the relationships with lower order polynomials. Specifically, we assume
\[
\lambda_{r1} = h_{\lambda}(x_r; \beta_{\lambda}) \quad \mu_r = h_{\mu}(x_r; \beta_{\mu}) \quad \ln \sigma^2_r = h_{\sigma}(x_r; \beta_{\sigma})
\]
and consider polynomial functions for \(h_{\cdot}(x_r; \cdot)\). Meanwhile, for random vector \((u_r, v_r, 1-u_r-v_r) \sim \text{Multinoulli}(\gamma_{ra}, \gamma_{rb}, 1-\gamma_{ra}-\gamma_{rb})\), we perform polynomial multinomial logistic regression, assuming
\[
E \left[ \ln \left( \frac{\gamma_{ra}}{1-\gamma_{ra}-\gamma_{rb}} \right) | x_r \right] = h_a(x_r; \beta_a) \quad \text{and} \quad E \left[ \ln \left( \frac{\gamma_{rb}}{1-\gamma_{ra}-\gamma_{rb}} \right) | x_r \right] = h_b(x_r; \beta_b)
\]
are polynomial functions.

The log likelihood of Eq. (3.2) still partitions into two parts: \(l_{in}(\beta_{\lambda} \mid y)\) involves parameters \(\beta_{\lambda} = (\beta_{\lambda}, \beta_{\mu}, \beta_{\sigma})\) and \(l_{out}(\beta_{\gamma} \mid u, v)\) involves parameters \(\beta_{\gamma} = (\beta_a, \beta_b)\). We maximize
\[
l_{in}(\beta_{\lambda} \mid y) = \sum_{r=1}^{R} u_r v_r \left\{ -\frac{1}{2} \ln(2\pi \exp[h_{\sigma}(x_r; \beta_{\sigma})]) - \frac{[g(y_r; h_{\lambda}(x_r; \beta_{\lambda}), \lambda_2) - h_{\mu}(x_r; \beta_{\mu})]^2}{2 \exp[h_{\sigma}(x_r; \beta_{\sigma})]} \right. \\
\left. + \ln \left| \frac{dg(y; h_{\lambda}(x_r; \beta_{\lambda}), \lambda_2)}{dy} \right|_{y=y_r} \right\} \quad (3.4)
\]
using the BFGS algorithm (Fletcher, 1987). We use \texttt{multinom()} in \texttt{R} package \texttt{nnet} to maximize
\[
l_{out}(\beta_{\gamma} \mid y) = \sum_{r=1}^{R} \left\{ (1-u_r)h_a(x_r; \beta_a) + (1-v_r)h_b(x_r; \beta_b) - \ln \left[ 1 + e^{h_a(x_r; \beta_a)} + e^{h_b(x_r; \beta_b)} \right] \right\},
\]
over \(\beta_{\gamma} = (\beta_a, \beta_b)\).
3.2.5 ROC curves

Receiver operating characteristic (ROC) curves are useful for comparing multiple implementations of a hypothesis test. In this case, we wish to use a \( z \) test for either null hypothesis \( H_0 : s \geq 0 \) or \( H_0 : s \leq 0 \), both one-sided tests, comparing multiple methods to transform the data. Because of the non-zero locus-dependent point masses, \( \gamma_{rA} \) and \( \gamma_{rB} \), it is not possible to display traditional ROC curves of the true and false positive rates as a function of a continuously varying threshold against which the test statistics \( z_r \) are compared. However, since the transformed, standardized variables are supposed to be standard normal, we can obtain an ROC curve by continuously varying significance level \( \alpha \) and comparing each \( z_r \) to the appropriate critical value at locus \( r \). Specifically, for a one-sided test of positive selection when \( \gamma_{rB} < \alpha \), locus \( r \) is considered a positive result at significance level \( \alpha \) if

\[
z_r = \frac{y_r - \mu_r}{\sigma_r} \geq q_{1-\alpha+\gamma_{rB}},
\]

where \( q_{1-\alpha+\gamma_{rB}} \) is the standard normal quantile. When \( \gamma_{rB} \geq \alpha \), then locus \( r \) cannot be evaluated and is removed from the set of tested loci. As a result, the estimated ROC curves may not be strictly increasing functions of the false positive rate because of sample variation. In addition, if the transformation has not achieved normality, the false positive rate (FPR) when significance level \( \alpha = 0.05 \) may not equal 0.05. If the achieved FPR > 0.05 when \( \alpha = 0.05 \), then the method is said to be liberal; otherwise the method is conservative.

3.3 Results

3.3.1 Normality

To assess whether any of the proposed transformations could achieve near-normality under varying simulation conditions, we sample \( n_s = 1,000 \) non-fixed allele relative frequency changes and estimate \( \hat{\gamma}_a, \hat{\gamma}_b, \hat{\lambda}_1, \hat{\mu}, \) and \( \hat{\sigma}^2 \) for each of the 6,000 conditions of Simulation A. Under the neutral Wright-Fisher model, an allele becomes fixed at time point \( i \) if \( x_{r,i} \in \{0, 1\} \), but we assume scaled changes of “non-fixed” alleles are transformable to normally distributed random variables. By sampling equal numbers of non-fixed observations per simulation condition, the goal is to achieve accurate estimates of parameters \( \lambda_1, \mu, \) and \( \sigma^2 \) for every condition, even when \( 1 - \gamma_a - \gamma_b \) is small. We obtain parameter estimates for each setting of \( (N, T, x) \) and each tested transformation, then transform the data, and evaluate the uncensored, transformed data for evidence of normality (Table 3.3). The original Wright-Fisher scaled increments are clearly not normal, with
Table 3.3: Average normality metrics for each $N$. Skewness, kurtosis, and Shapiro-Wilk log $p$-values computed on datasets of size $n_s = 1,000$ before and after transformation are averaged across choices of initial allele frequency $x$ and elapsed generations $T$. Averages under $N = 200$ are calculated from up to 1,000 $(x, T)$ combinations, while the $N = 1,000$ results are comprised of up to 5,000 values of $(x, T)$. Averages over subsets of the data (visualized in Fig. 3.3) where boundary point masses, $\gamma_A$ or $\gamma_B$, are below 0.05 are also reported. Transformations are abbreviated as in Table 3.1.

<table>
<thead>
<tr>
<th>$N$</th>
<th>Subset</th>
<th>Metrics</th>
<th>WF</th>
<th>BC</th>
<th>YJ</th>
<th>MAN</th>
<th>IHS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Skewness</td>
<td>−0.507</td>
<td>0.148</td>
<td>0.000</td>
<td>−0.091</td>
<td>0.084</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kurtosis</td>
<td>2.629</td>
<td>2.247</td>
<td>2.032</td>
<td>1.998</td>
<td>2.069</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shapiro Wilk log $p$</td>
<td>−38.936</td>
<td>−21.120</td>
<td>−27.644</td>
<td>−30.187</td>
<td>−29.934</td>
</tr>
<tr>
<td>200</td>
<td>$\gamma_A &lt; 0.05$</td>
<td>Skewness</td>
<td>−0.609</td>
<td>0.123</td>
<td>−0.017</td>
<td>−0.108</td>
<td>0.134</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kurtosis</td>
<td>2.836</td>
<td>2.308</td>
<td>2.078</td>
<td>2.041</td>
<td>2.127</td>
</tr>
<tr>
<td></td>
<td>$\gamma_B &lt; 0.05$</td>
<td>Skewness</td>
<td>−0.301</td>
<td>0.085</td>
<td>0.011</td>
<td>−0.029</td>
<td>−0.015</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kurtosis</td>
<td>2.683</td>
<td>2.575</td>
<td>2.437</td>
<td>2.404</td>
<td>2.392</td>
</tr>
<tr>
<td>1,000</td>
<td>$\gamma_A &lt; 0.05$</td>
<td>Skewness</td>
<td>−0.514</td>
<td>0.150</td>
<td>−0.003</td>
<td>−0.093</td>
<td>0.074</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kurtosis</td>
<td>2.668</td>
<td>2.286</td>
<td>2.049</td>
<td>2.016</td>
<td>2.077</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shapiro Wilk log $p$</td>
<td>−38.518</td>
<td>−19.488</td>
<td>−27.003</td>
<td>−29.568</td>
<td>−29.220</td>
</tr>
<tr>
<td></td>
<td>$\gamma_B &lt; 0.05$</td>
<td>Skewness</td>
<td>−0.607</td>
<td>0.126</td>
<td>−0.018</td>
<td>−0.108</td>
<td>0.118</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kurtosis</td>
<td>2.863</td>
<td>2.347</td>
<td>2.095</td>
<td>2.059</td>
<td>2.131</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shapiro Wilk log $p$</td>
<td>−38.905</td>
<td>−16.245</td>
<td>−24.957</td>
<td>−27.955</td>
<td>−27.369</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Skewness</td>
<td>−0.305</td>
<td>0.079</td>
<td>0.006</td>
<td>−0.030</td>
<td>−0.017</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kurtosis</td>
<td>2.725</td>
<td>2.608</td>
<td>2.468</td>
<td>2.437</td>
<td>2.425</td>
</tr>
</tbody>
</table>
negative skew and negative excess kurtosis (platykurtosis). Overall, the transformations improve normality, as measured by the Shapiro-Wilk test (Shapiro and Wilk, 1965). The transformed data show reduced skewness, but slightly increased average platykurtosis, probably because of censoring in the right tail. A more detailed view across simulation conditions (Fig. 3.2) shows the transformations greatly reduce the absolute skewness observed for high initial allele frequencies and moderate kurtosis everywhere. The Shapiro-Wilk test reveals substantial improvement in normality as measured by an increase in the test \( p \)-values for all transformations, but particularly the Box-Cox transformation, trailed by YJ, then IHS and Manly in indeterminate order (Fig. 3.3). Since over 98\% of the Shapiro-Wilk tests of untransformed Wright-Fisher data have \( p \)-values < 0.05, increases in the \( p \)-values represent improvements in normality. In fact, \( p \)-values for untransformed data exceed 0.05 for only 15 of the \( N = 200 \) simulations, all with \( x < 0.65 \) and \( T/N < 0.1 \), and only 116 of the \( N = 1,000 \) simulations, all with \( x \leq 0.8 \) and \( T/N \leq 0.12 \), i.e. in the lower left extreme corners of the images in Figs. 3.3(a)–(b). All transformations show the greatest improvement in normality for high initial allele frequencies \( x \) and intermediate time lapses, when the untransformed data are least normal. The cumulative distribution functions shown for four choices of initial allele frequency \( x \) at each effective population size \( N \) [Figs. 3.3(c)–(d)], confirm the left skew and relative improvement in normality of the four transformations.

The FITR test cannot test neutrality of a single locus when there is substantial chance of allele fixation, i.e. when the point masses, \( \gamma_A \) and \( \gamma_B \), are large. The useful region for two-tailed hypothesis tests with significance level not exceeding 0.05 is below the blue, solid line in Figs. 3.3(a)–(b): above this line, \( \gamma_B > 0.025 \). Below the black solid line is the region where one-sided tests of neutrality with alternative hypothesis \( H_1 : s > 0 \) can reject at \( \alpha \leq 0.05 \); below the black dashed line, one-sided tests of neutrality with \( H_1 : s < 0 \) can reject at \( \alpha \leq 0.05 \). Here \( s \) is the selection coefficient, and \( s > 0 \) indicates the allele relative frequency will tend to increase because of selection. In all relevant regions, Box-Cox achieves the best normality, but all transformations improve normality relative to the untransformed Wright-Fisher data (Table 3.3). Outside these regions, a collection of independent sites may be tested for an unusual pattern of change. For example, one could test if the observed proportion of \( y > c \) differs from the expected proportion \( \gamma_B + 1 - \Phi(c) \) for any \( c \) if the normality of the transformed \( y \) can be trusted.

### 3.3.2 Modeling \( \lambda_{r1}, \mu_r, \sigma_r, \gamma_{ra} \) and \( \gamma_{rb} \) as functions of \( x_r \)

The optimal transformation, and hence \( \lambda_{r1}, \mu_r, \sigma_r, \gamma_{ra} \) and \( \gamma_{rb} \), depend on the initial allele frequency \( x_r \), the effective population size \( N \), and the elapsed time \( \Delta \tau \). While \( N \) and \( \Delta \tau \) are constant across reference loci,
Table 3.4: Normality of transformed data when $\lambda_{r1}, \mu_r$, and $\log \sigma_r^2$ are polynomial functions of initial allele frequency $x_r$. Data are from simulations B1.1 and B2.1 with $s = 0$.

<table>
<thead>
<tr>
<th>$N$</th>
<th>Metrics</th>
<th>WF</th>
<th>BC</th>
<th>YJ</th>
<th>MAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>Polynomial degree</td>
<td>(0, 0, 0)</td>
<td>(1, 3, 3)</td>
<td>(1, 3, 3)</td>
<td>(1, 3, 3)</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>-2.640</td>
<td>0.199</td>
<td>0.102</td>
<td>-0.227</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>15.992</td>
<td>2.446</td>
<td>2.259</td>
<td>2.288</td>
</tr>
<tr>
<td></td>
<td>Shapiro Wilk log $p$</td>
<td>-125.390</td>
<td>-32.147</td>
<td>-34.500</td>
<td>-60.576</td>
</tr>
<tr>
<td>1000</td>
<td>Polynomial degree</td>
<td>(0, 0, 0)</td>
<td>(1, 3, 2)</td>
<td>(1, 2, 2)</td>
<td>(2, 3, 2)</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>-2.737</td>
<td>0.112</td>
<td>0.014</td>
<td>-0.068</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>23.550</td>
<td>2.596</td>
<td>2.215</td>
<td>2.175</td>
</tr>
<tr>
<td></td>
<td>Shapiro Wilk log $p$</td>
<td>-129.923</td>
<td>-25.583</td>
<td>-46.329</td>
<td>-50.326</td>
</tr>
</tbody>
</table>

it is not feasible to sample $R$ reference loci with the exact same initial allele frequencies $x_r$ in practice. Thus, we must account for a relationship between $x_r$ and these parameters. For a given combination $(N, \Delta r)$, we choose to model $\lambda_{r1}, \mu_r$, and $\ln \sigma_r^2$ as low order polynomial functions of $x_r$. Similarly, we assume the counts $\sum_{r=1}^{R} (1 - u_r)$ and $\sum_{r=1}^{R} (1 - v_r)$ of alleles striking each boundary satisfy a multinomial logistic regression on powers of $x_r$.

To demonstrate that such polynomial regression is adequate to explain the dependence on $x_r$, we perform polynomial regression with the parameters $\hat{\lambda}_{r1}, \hat{\mu}_r$, and $\ln \hat{\sigma}_r^2$ estimated from 1,000 uncensored, simulated observations at relative initial allele frequency $x_r$ for a fixed $N$ and $T$. Fig. 3.4 shows cubic fits are largely adequate for explaining the nonlinear dependence of each of $\lambda_{r1}, \mu_r$, and $\ln \sigma_r^2$ on $x_r$ for both $N \in \{200, 1000\}$ and a variety of elapsed generations $T$. Similarly, Fig 3.5 demonstrates that quartic polynomials are adequate to explain the dependence of both $\gamma_{rn}$ and $\gamma_{rb}$ on $x_r$. Since data will typically consist of far fewer observations per $x_r$, particularly large $x_r$, we expect lower order polynomials to suffice in practice.

### 3.3.3 Transforming datasets with mixed initial allele frequencies $x_r$

In practice, data will consist of pairs $(x_r, y_r)$ for each reference locus $r$, where relative initial allele frequency $x_r$ varies. Without loss of generality, $x_r \geq 0.5$ if we consider only biallelic loci and track the allele that was dominant at the initial time point. To allow the transformation to vary as a function of $x_r$, we introduce the polynomial dependence on $x_r$ into the log likelihood Eq. (3.2) and maximize over coefficients $\beta_N$ (see Methods). To choose an appropriate polynomial order, we fit up to order two polynomials for all of $\lambda_1, \mu$ and $\sigma^2$, then choose the best model by minimizing AIC. For the multinomial regression we fit increasing polynomial orders and choose the order yielding the first minimum AIC. The cdfs in Fig. 3.6 demonstrate...
that the estimated transformation achieves good normality compared to the original, untransformed Wright-Fisher data, and comparable to the performance achieved by the transformations estimated from the excess data available in Simulation A.

### 3.3.4 Detecting selection

We also simulated non-neutral loci to examine method performance in the presence of selection (Simulation B, 3.2). To better compare the tests, we prepare ROC curves by varying the significance from 0 to 1 and computing the true positive and false positive rates for each transformation or the Wright-Fisher data. At each significance level, there will be a fraction of loci that cannot be tested. For example, the number of positive cases cannot be determined when testing $H_1 : s > 0$ at significance level $\alpha = 0.05$ for any loci with $x_0$ large enough to put $\gamma_B \geq 0.05$. Among the remaining loci, we compute the positive results as those exceeding their respective critical value $q_{1-\alpha+\gamma_B}$ from the standard normal distribution. A fraction of these are true positives with $s > 0$. Fig. 3.7 shows the ROC curves for simulations B1.2, B1.3, B1.5, and B1.6. The Box-Cox transformed data clearly have the best performance as a test of $H_0 : s \leq 0$ when true $s > 0$. The performance considerably exceeds the untransformed data, while the YJ and Manly transformations achieve intermediate performance. When $H_0 : s \geq 0$ and the true $s < 0$, there is little difference in the ROC curves, but the FPR better matches the specified significance level $\alpha$, especially for YJ and Manly transformations. The untransformed data are overly conservative, and the Box-Cox transformation is slightly over-conservative, while the remaining transformations achieve the desired FPR.

The numerical false positive rates achieved for various choices of selection coefficient $s$, significance level $\alpha$ are shown for simulation conditions B1 and B2 in Table 3.5. All transformation methods vye for best control of FPR, but are generally conservative for $\alpha \leq 0.05$. Similarly, the achieved numerical power are shown for simulation condition B1 (Table 3.6) and B2 (Table 3.7). Box-Cox has the best power to detect alternative $H_1 : s > 0$ on a major allele ($x_0 > 0.5$), but Manly has the best power to detect $H_1 : s < 0$, although Box-Cox is not far behind. Overall, Box-Cox seems to achieve the best performance.

### 3.4 Discussion

We have proposed an improvement to the FITR test for selection, which substantially improves performance of this test by insuring better fidelity to the normality assumption. The original authors of FITR (Nishino, 2013) presumed that the normal approximation would hold in conditions where the test...
Table 3.5: False positive rates at significance levels $\alpha \in \{0.001, 0.01, 0.025, 0.05, 0.1\}$ for data simulated under Simulation B1.1 or B2.1 (see Table 3.2). Transformations estimated from 1,000 simulated observations; false positive rates estimated from 10,000 simulated observations, including 1,000 used in estimation. The transformation achieving FPR closest to the significance level is bolded.

<table>
<thead>
<tr>
<th>$H_1: s &lt; 0$</th>
<th>$H_1: s &gt; 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>$\psi$</td>
<td>$\psi$</td>
</tr>
<tr>
<td>WF</td>
<td>BC</td>
</tr>
<tr>
<td>0.001</td>
<td>0.000</td>
</tr>
<tr>
<td>0.010</td>
<td>0.000</td>
</tr>
<tr>
<td>0.025</td>
<td>0.000</td>
</tr>
<tr>
<td>0.050</td>
<td>0.000</td>
</tr>
<tr>
<td>0.100</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Simulation B1.1 ($N_e = 200, T = 40, s = 0$)
Simulation B2.1 ($N_e = 1,000, T = 200, s = 0$)

WF BC MAN YJ
Table 3.6: True positive rates for varying selection intensities for data simulated under Simulation N = 200, T = 40 (see Table 3.2). The transformation with maximum power is bolded for each selection intensity and significance level. See Table 3.5 for further details.

<table>
<thead>
<tr>
<th>s</th>
<th>α</th>
<th>H1 : s &gt; 0</th>
<th>H1 : s &lt; 0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>WF</td>
<td>BC</td>
</tr>
<tr>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.01</td>
<td>0.000</td>
<td>0.010</td>
<td>0.001</td>
</tr>
<tr>
<td>0.025</td>
<td>0.000</td>
<td>0.023</td>
<td>0.017</td>
</tr>
<tr>
<td>0.05</td>
<td>0.000</td>
<td>0.058</td>
<td>0.043</td>
</tr>
<tr>
<td>0.1</td>
<td>0.000</td>
<td>0.117</td>
<td>0.105</td>
</tr>
<tr>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.01</td>
<td>0.000</td>
<td>0.013</td>
<td>0.002</td>
</tr>
<tr>
<td>0.025</td>
<td>0.000</td>
<td>0.053</td>
<td>0.034</td>
</tr>
<tr>
<td>0.05</td>
<td>0.000</td>
<td>0.107</td>
<td>0.087</td>
</tr>
<tr>
<td>0.1</td>
<td>0.000</td>
<td>0.193</td>
<td>0.183</td>
</tr>
<tr>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.01</td>
<td>0.000</td>
<td>0.147</td>
<td>0.038</td>
</tr>
<tr>
<td>0.05</td>
<td>0.000</td>
<td>0.342</td>
<td>0.241</td>
</tr>
<tr>
<td>0.1</td>
<td>0.000</td>
<td>0.509</td>
<td>0.420</td>
</tr>
<tr>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.01</td>
<td>0.000</td>
<td>0.488</td>
<td>0.173</td>
</tr>
<tr>
<td>0.05</td>
<td>0.000</td>
<td>0.722</td>
<td>0.521</td>
</tr>
<tr>
<td>0.1</td>
<td>0.000</td>
<td>0.862</td>
<td>0.703</td>
</tr>
<tr>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.01</td>
<td>0.000</td>
<td>0.008</td>
<td>0.001</td>
</tr>
<tr>
<td>-0.001</td>
<td>0.025</td>
<td>0.000</td>
<td>0.024</td>
</tr>
<tr>
<td>0.05</td>
<td>0.000</td>
<td>0.055</td>
<td>0.043</td>
</tr>
<tr>
<td>0.1</td>
<td>0.000</td>
<td>0.114</td>
<td>0.107</td>
</tr>
<tr>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.01</td>
<td>0.000</td>
<td>0.005</td>
<td>0.001</td>
</tr>
<tr>
<td>-0.01</td>
<td>0.025</td>
<td>0.000</td>
<td>0.010</td>
</tr>
<tr>
<td>0.05</td>
<td>0.000</td>
<td>0.025</td>
<td>0.019</td>
</tr>
<tr>
<td>0.1</td>
<td>0.000</td>
<td>0.114</td>
<td>0.107</td>
</tr>
<tr>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.01</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>-0.05</td>
<td>0.025</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.05</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.1</td>
<td>0.000</td>
<td>0.002</td>
<td>0.001</td>
</tr>
<tr>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.01</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>-0.1</td>
<td>0.025</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.05</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.1</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Table 3.7: True positive rates for varying selection intensities given data simulated under Simulation B2 ($N = 1,000$, $T = 200$) (see Table 3.2). The transformation with maximum power is bolded for each selection intensity and significance level. See Table 3.5 for further details.

<table>
<thead>
<tr>
<th>$s$</th>
<th>$H_1 : s &gt; 0$</th>
<th>$H_1 : s &lt; 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.001</td>
<td>NA 0.148 0.58 0.083</td>
<td>0.016 0.012 0.000 0.000</td>
</tr>
<tr>
<td>0.01</td>
<td>0.000 0.238 0.184 0.201</td>
<td>0.033 0.047 0.019 0.034</td>
</tr>
<tr>
<td>0.025</td>
<td>0.000 0.301 0.253 0.264</td>
<td>0.064 0.120 0.134 0.125</td>
</tr>
<tr>
<td>0.05</td>
<td>0.000 0.367 0.339 0.349</td>
<td>0.095 0.177 0.192 0.187</td>
</tr>
<tr>
<td>0.1</td>
<td>0.000 0.436 0.409 0.420</td>
<td>0.163 0.124 0.133 0.128</td>
</tr>
<tr>
<td>0.01</td>
<td>0.000 0.287 0.224 0.244</td>
<td>0.029 0.050 0.051 0.047</td>
</tr>
<tr>
<td>0.05</td>
<td>0.000 0.371 0.308 0.321</td>
<td>0.041 0.080 0.087 0.082</td>
</tr>
<tr>
<td>0.1</td>
<td>0.000 0.443 0.409 0.420</td>
<td>0.063 0.124 0.133 0.128</td>
</tr>
<tr>
<td>0.01</td>
<td>0.000 0.360 0.135 0.198</td>
<td>0.021 0.029 0.008 0.017</td>
</tr>
<tr>
<td>0.05</td>
<td>0.000 0.560 0.430 0.464</td>
<td>0.003 0.005 0.004 0.003</td>
</tr>
<tr>
<td>0.1</td>
<td>0.000 0.675 0.555 0.582</td>
<td>0.005 0.009 0.009 0.007</td>
</tr>
<tr>
<td>0.01</td>
<td>0.000 0.752 0.682 0.701</td>
<td>0.009 0.017 0.018 0.015</td>
</tr>
<tr>
<td>0.05</td>
<td>0.000 0.571 0.242 0.326</td>
<td>0.002 0.002 0.000 0.001</td>
</tr>
<tr>
<td>0.1</td>
<td>0.000 0.793 0.608 0.653</td>
<td>0.000 0.000 0.000 0.000</td>
</tr>
<tr>
<td>0.01</td>
<td>0.000 0.879 0.718 0.747</td>
<td>0.000 0.001 0.000 0.000</td>
</tr>
<tr>
<td>0.05</td>
<td>0.000 0.929 0.821 0.850</td>
<td>0.001 0.001 0.001 0.001</td>
</tr>
<tr>
<td>0.1</td>
<td>0.000 0.151 0.054 0.079</td>
<td>0.034 0.048 0.020 0.034</td>
</tr>
<tr>
<td>-0.001</td>
<td>0.000 0.227 0.177 0.192</td>
<td>0.047 0.084 0.090 0.084</td>
</tr>
<tr>
<td>0.05</td>
<td>0.000 0.283 0.239 0.249</td>
<td>0.064 0.125 0.136 0.129</td>
</tr>
<tr>
<td>0.1</td>
<td>0.000 0.347 0.321 0.329</td>
<td>0.097 0.177 0.192 0.186</td>
</tr>
<tr>
<td>0.01</td>
<td>0.000 0.122 0.046 0.068</td>
<td>0.048 0.071 0.033 0.056</td>
</tr>
<tr>
<td>-0.01</td>
<td>0.000 0.188 0.147 0.156</td>
<td>0.065 0.122 0.132 0.124</td>
</tr>
<tr>
<td>0.05</td>
<td>0.000 0.233 0.195 0.203</td>
<td>0.088 0.178 0.197 0.189</td>
</tr>
<tr>
<td>0.1</td>
<td>0.000 0.283 0.264 0.271</td>
<td>0.133 0.255 0.272 0.266</td>
</tr>
<tr>
<td>0.01</td>
<td>0.000 0.046 0.021 0.029</td>
<td>0.141 0.265 0.228 0.269</td>
</tr>
<tr>
<td>-0.05</td>
<td>0.000 0.075 0.057 0.060</td>
<td>0.191 0.418 0.450 0.438</td>
</tr>
<tr>
<td>0.05</td>
<td>0.000 0.088 0.072 0.075</td>
<td>0.254 0.535 0.563 0.552</td>
</tr>
<tr>
<td>0.1</td>
<td>0.000 0.108 0.096 0.098</td>
<td>0.366 0.637 0.653 0.649</td>
</tr>
<tr>
<td>0.01</td>
<td>0.000 0.017 0.007 0.010</td>
<td>0.244 0.574 0.565 0.585</td>
</tr>
<tr>
<td>-0.1</td>
<td>0.000 0.024 0.020 0.020</td>
<td>0.342 0.817 0.840 0.832</td>
</tr>
<tr>
<td>0.05</td>
<td>0.000 0.029 0.023 0.024</td>
<td>0.446 0.889 0.896 0.894</td>
</tr>
<tr>
<td>0.1</td>
<td>0.000 0.032 0.029 0.029</td>
<td>0.634 0.920 0.922 0.921</td>
</tr>
</tbody>
</table>
would be applied. In essence, users were assumed to avoid long time increments, high initial major relative allele frequencies, and small effective population size $N$. However, we have shown that the usefulness of the test can be extended into these boundaries by achieving normality through transformation. Beyond these boundaries, the test is not applicable because of large point masses in the boundary that exceed the significance level.

In our simulation, we sampled initial relative allele frequencies for the reference loci uniformly from the interval $[0.5, 1)$. After $T$ generations, loci with $x_r$ nearer to 0.5 will be overrepresented among the unfixed loci used to estimate the transformation. We consider such bias a natural phenomena and do not compensate for it, but it implies that the transformation is estimated with greater uncertainty for loci starting with large $x_r$. If reference loci were abundant, one could imagine a strategy to over-sample reference loci with large $x_r$, though it would require knowledge of $N$, $T$ and the unknown $\gamma_A$ and $\gamma_B$ to choose a sampling scheme that would end with a balanced unfixed sample. In addition, biological systems will have different distributions of $x_r$ at neutral loci depending on their demographic and evolutionary history. It is not clear how such distributions will affect the method, but as long as their sufficient independent reference loci and selected loci have similar initial allele frequencies, we expect the method to adapt well.

While we only consider one-sided tests, $H_0 : s \leq 0$ to detect positive selection and $H_0 : s \geq 0$ to detect negative selection, it is natural extension to consider two-tailed tests. On the other hand, because of the relative inequality of the point masses in each tail, especially when non-normality is pervasive, it may be desirable to only perform one tail tests based on the observed $x_r$, simply sacrificing the ability to detect positive selection of dominant alleles and negative selection of minor alleles. We have left these choices to the user and focused on the performance of our method in one-tailed tests.

We have evidence (data not shown) that the Box-Cox transformation does not approximate the true Wright-Fisher distribution as well as a recently proposed Beta-with-spikes approximation (Tataru et al., 2015). Thus, one could imagine that a Likelihood Ratio Test where the Beta-with-spikes is used to estimate effective population size $N$, $\Delta \tau$, and $s$ may perform better. On the other hand, there is very little information about both $N$ and $\Delta \tau$ separately in relative allele frequency increments (Tataru et al. (2015) actually fixed $\Delta \tau$ arbitrarily and just estimated $N$). Since Tataru et al. (2015) focused on estimation of $N$ and did not use their method to test for selection, it is an obvious future direction of research to compare these two methods for testing selection.
Figure 3.2: Skewness and kurtosis for untransformed simulated data under $N \in \{200, 1000\}$ and various transformations of the same data. Each top plot is comprised of 1,000 grid points, one for each simulation condition under $N = 200$, with darker red hues representing more favorable outcomes and, in contrast, darker blue hues representing less favorable outcomes. Along the $x$-axis is the initial allele frequency and along the $y$-axis is the number of generations $T$, scaled by effective population size $N$. The bottom plots are comprised of 5,000 grid points for the simulation conditions under $N = 1,000$. Transformations are Box-Cox (BC), Yeo-Johnson (YJ), Manly (MAN), and Inverse Hyperbolic Sine (IHS) [see Table 3.1].
Figure 3.3: Shapiro-Wilk test of normality for all simulation conditions. Each image is comprised of (a) 1,000 or (b) 5,000 grid points, one for each simulation condition \((x, T)\), and the value displayed in each grid is the difference in log \(p\)-values of the Shapiro-Wilk test for normality applied to all \(n_s\) simulated observations from the corresponding \((x_i, T)\). The log \(p\)-value for transformed data is subtracted from log \(p\)-value for untransformed Wright-Fisher data, so large numbers, deep blue, indicate favorable performance of the respective transformation; small numbers, deep red, indicate no difference in the \(p\)-value. The black solid line delineates the boundary between conditions where \(\gamma_B > 0.05\) (above) and \(\gamma_B < 0.05\) (below). The black dashed line similarly delineates the \(\gamma_A = 0.5\) boundary with \(\gamma_A > 0.5\) above the line. Blue versions of these lines indicate the boundary \(\gamma_A = 0.025\) or \(\gamma_B = 0.025\). Cumulative distribution functions for starred settings in (a) [respectively (b)] are displayed in (c) [respectively, (d)]. Transformations abbreviated as in Table 3.1.
Figure 3.4: Polynomial fits of (a) $\lambda_{r1}$, (b) $\mu_r$, and (c) $\sigma_r^2$ at the indicated scaled time lapse $T/N$ in each facet and $N = 200$ (left) or $N = 1,000$ (right). Points are initial allele relative frequencies $x_r$ plotted against maximum likelihood estimates $\hat{\lambda}_{r1}$, $\hat{\mu}_r$, and $\hat{\sigma}_r^2$ obtained by maximizing the likelihood Eq. (3.2) for $n_s = 1,000$ simulated $y_r^{(1)}, y_r^{(2)}, \ldots, y_r^{(1000)}$ observed within $(a_r, b_r)$. The colors indicate the transformation, the line is the predicted cubic polynomial curve, and the gray band represents the 95% confidence interval.
Figure 3.5: Multinomial logistic regression fits of $\gamma_A$ and $\gamma_B$ at the indicated scaled time lapse $T/N$ in each facet and $N = 200$ (top) or $N = 1,000$ (bottom). Counts of simulated $y_r$ fixing on the left $(a_r)$ and right $(b_r)$ boundary when a total of $n_s = 1,000$ non-boundary $y_r \in (a_r, b_r)$ were fitted by multinomial logistic regression on powers of $x_r$ for each choice of $N$ and $T$. The line is the predicted mean from a quartic polynomial of $x_r$, which was necessary for the fit at $T/N = 0.2$, but excessive for other time points. A gray band showing 95% confidence intervals is obscured by the observations.
Figure 3.6: Empirical cumulative distribution functions for (a) $N = 200, T = 40, s = 0$ and (b) $N = 1,000, T = 200$. Transformations are abbreviated as in Table 3.2, additionally WF = Wright-Fisher scaled data $y_t$, WFz = standardized Wright-Fisher scaled data, SimA-BC = Box-Cox transformed data using $\lambda_{r1}, \mu_r$, and $\sigma_r^2$ estimated from simulation A data.
Figure 3.7: ROC curves for $N = 200$, $T = 40$ data on the left and $N = 1,000$, $T = 200$ data on the right. Plots (a) and (b) show curves for testing the one-sided hypothesis $H_0 : s \leq 0$ against $H_1 : s > 0$. Plots (c) and (d) show curves for testing the one-sided hypothesis $H_0 : s \geq 0$ against $H_1 : s < 0$. Vertical lines demarcate the classic $\alpha = 0.05$ significance level. Dots indicate the point where the values of each respective transformation cross the 0.05 standard normal quantile. If normality has been achieved, then the dots should appear when the ROC curve crosses the vertical line.
CHAPTER 4. CONCLUSION

4.1 Future Work

Highlighted in the following sections are topics that due to the varied limitations of this research were not explored to full potential. These research areas are left as future work and the broader impact that any of these potential areas, individually or collectively, has on the conclusion of this dissertation is not known. Lastly, we briefly offer initial strategies to how one may address these problems in the future.

4.1.1 Conserving Normal Quantiles in the Tails

Within the application of the normalizing transformations, there is evidence that the transformations do not always conserve the tail probabilities of the normal distribution. When comparing quantiles of the post-transformed data to those of the standard normal distribution, often a set of quantiles in one tail may not adequately resemble those of the normal distribution. Harrell-Davis estimates of several quantiles from pre- and post-transformed data display the issues with fitting both tails simultaneously. As seen in Fig. 4.1, the upper tail quantiles under the transformations consistently underestimate the quantiles of the normal distribution, while the lower tail estimates are more stable and closer to the truth.

Since the utility of the FITR test relies heavily on the assumption of normality, even subtle inconsistencies between the transformed tail quantiles and normal quantiles can have dramatic effects on the outcomes of the FITR test. As a result, in chapter 3 we limited the tests of selection to one-tailed test, but we would like to extend the usefulness of the FITR to included two-tailed tests for selection. However, in order to properly test all possible alternative hypotheses of the FITR as well as have higher power and better control of the type-1 error rate, the tail distributions post-transformation tails must be adequate approximations to the normal distribution. Taking a non-parametric approach to this problem, one could employ the normal quantile transformation (Cer et al., 2014) to the original data instead of the transformations found in Table 3.1. This rank-based technique has proven to be more robust and less biased in comparison to other normalizing techniques (Cer et al., 2014). A second, slightly more complicated option would partition the data into two disjoint subsets and apply a normalizing transformation for each subset individually. The premise is to
transform the tails separately, since the current transformations are not simultaneously successful in both tails.

### 4.1.2 The Challenge of Linkage Disequilibrium

The fundamental requirement of the FITR test is a set of unlinked and neutrally evolving reference loci. Fortunately, Nishino (2013) found that selection at the reference loci had negligible effect on type I error so long as $s < 1$. To apply FITR to our data, we assume that most segregating loci are neutral or only weakly ($s < 1$) selected, so that most are qualified reference loci. However, in practice it is nearly impossible for such a strong assumption of independence to hold due to linkage disequilibrium. This is particularly true for viruses, such as PRRSV and other organisms with short genomes, where the distance between two loci may not be large enough to overcome the spatial association. The nonrandom association of alleles at multiple loci, can intensify or dwarf the effects of selection of linked loci (Slatkin, 2008). For instance, tightly linked loci allow genetic hitchhiking when one of the associated loci is being selected (Illingworth and Mustonen, 2011). Although temporal data offer powerful insight into the evolutionary impact of natural selection among linked sites (Illingworth and Mustonen, 2011, 2012; Turner et al., 2011; Kofler and Schlötterer, 2014; Terhorst et al., 2015), the issue of linkage disequilibrium still remains a complex problem (Illingworth and Mustonen, 2011, 2012).

### 4.1.3 Alternative Models of the Distribution of Allele Frequency (DAF)

The distribution of allele frequency (DAF) models how population allele frequencies change over time, i.e. $f(x_j | x_{j-1})$. In the WF-HMM, the DAF is represented by the transition matrix of the allele frequencies. However, when $N$ is large, the solutions to the DAF are no longer available in closed form. For this reason, several existing methods have utilized the continuous distributions to approximate the WF-HMM DAF. Since allele frequencies lie within the unit interval, the beta distribution (Gautier and Vitalis, 2013; Sirén et al., 2011; Tataru et al., 2015; Hui and Burt, 2015) is a natural choice as the distributional form of the DAF. However, the truncated normal distribution has also been implemented in recent analysis (Nicholson et al., 2002; Gautier et al., 2010). Despite the mathematical convenience of the continuous approximation, neither the truncated normal nor beta distribution account for the positive probability that can accumulate at the boundaries $\{0, 1\}$. In the presence of small effective population sizes or long time intervals, the mass at these atoms can no longer be considered negligible.
As a remedy, Tataru et al. (2015, 2016) propose to model the DAF using a beta with spikes (BWS) approximation. The refined shape parameters, $a_t$ and $b_t$, of the beta distribution are functions of the conditional mean and variance after $t$ generations of the latter allele frequency falling inside the open unit interval. The two point masses, $\gamma_{A,t}$ and $\gamma_{B,t}$, capture the probability of the allele frequency after $t$ generations on reaching extinction or fixation, respectively. Then, the parameters of the beta with spikes, $\theta = \{a_t, b_t, \gamma_{A,t}, \gamma_{B,t}\}$, are defined recursively as

$$a_t = \left( \frac{E[X_t \mid X_0, X_t \notin \{0, 1\}](1 - E[X_t \mid X_0, X_t \notin \{0, 1\}])}{\text{Var}(X_t \mid X_0, X_t \notin \{0, 1\})} - 1 \right) E[X_t \mid X_0, X_t \notin \{0, 1\}]$$

$$b_t = \left( \frac{E[X_t \mid X_0, X_t \notin \{0, 1\}](1 - E[X_t \mid X_0, X_t \notin \{0, 1\}])}{\text{Var}(X_t \mid X_0, X_t \notin \{0, 1\})} - 1 \right) (1 - E[X_t \mid X_0, X_t \notin \{0, 1\}])$$

$$\gamma_{A,t} = \gamma_{A,t-1} + (1 - \gamma_{A,t-1} - \gamma_{B,t-1}) \frac{B(a_t, b_t + N)}{B(a_t, b_t)}$$

$$\gamma_{B,t} = \gamma_{B,t-1} + (1 - \gamma_{A,t-1} - \gamma_{B,t-1}) \frac{B(a_t + N, b_t)}{B(a_t, b_t)}$$

where

$$E[X_t \mid X_0, X_t \notin \{0, 1\}] = \frac{x_t - \gamma_{B,t}}{1 - \gamma_{A,t} - \gamma_{B,t}}$$

$$\text{Var}(X_t \mid X_0, X_t \notin \{0, 1\}) = \frac{x_t(1 - x_t) \left[ 1 - \left( 1 - \frac{1}{N} \right)^t \right] + x_t^2 - \gamma_{B,t}}{1 - \gamma_{A,t} - \gamma_{B,t}} - \left( E[X_t \mid X_0, X_t \notin \{0, 1\}] \right)^2,$$

initialized from $\gamma_{A,0} = \gamma_{B,0} = 0$ and $B(a, b)$ is the beta function.

Tataru et al. (2016) showed, by means of the Hellinger distance (Hellinger, 1909), that the BWS more closely approximates the true WF-HMM DAF than its truncated normal and traditional beta counterparts. However, $\hat{\gamma}_{A,t}$ and $\hat{\gamma}_{B,t}$ were shown to poorly estimate the extinction and fixation probabilities when the initial allele frequencies are close to the boundary (Tataru et al., 2015). Instead, the point estimates of the $\gamma_{A,0}$ and $\gamma_{B,0}$ from simulation A (see Table 3.2) were more accurate than the $\hat{\gamma}_{A,t}$ and $\hat{\gamma}_{B,t}$ from BWS. We obtained the Beta with Hellams spikes (BWHS) approximation to the DAF by replacing the $\hat{\gamma}_{A,t}$ and $\hat{\gamma}_{B,t}$ from BWS with the estimates from simulation A. The Hellinger distance between the true WF-HMM DAF and truncated normal, BWHS, BWS, and truncated normal after Box-Cox transformation for various $x_0, T$ and $N$ combinations are shown in Fig. 4.2. The BWHS best resembled the true WF-HMM DAF. Interestingly, the Box-Cox transformed data appeared to have superior performance over a small region where BWS and BWHS most deviate from the true DAF. The latter observation suggests that the Box-Cox transformation may be superior to all other methods for certain choices of $x_0$ and $T$, especially for large $N_c$.

The preceding analysis relied on $\gamma_{A,0}$ and $\gamma_{B,0}$ estimated from a large numbers of simulations at each $x_0, T$ and $N$ combination displayed in Fig. 4.2. In real data, the true $x_0$, $T$, and $N$ will not be known. However,
since Fig. 3.5 shows that $x_0$ and $T$ are good predictors of the extinction and fixation probabilities when $N$ is constant, we propose to improve the accuracy of the BWS model by replacing $\hat{A}_t$ and $\hat{B}_t$ with estimates of the spikes from a multinomial logistic regression (BWS-MLR) where $x_0$ and $T$ are model covariates (see §3.2.4).

### 4.1.3.1 Estimating $N$

The efficacy of the selective force is affected by population size. Variation in small populations is more likely to overwhelm the selective force, whereas the same selective force will have a more pronounced impact relative to genetic drift within a larger population. Thus, modeling assumptions regarding population size have major implications on the utility of selection detection methods. Several approaches for detecting selection that are independent of the nuisance parameter $N$ (Feder et al., 2014; Nishino, 2013) have been developed, while others estimate $N$ directly within the WF-HMM framework (Williamson and Slatkin, 1999; Anderson et al., 2000; Bollback et al., 2008; Hui and Burt, 2015). Assuming the initial allele frequency $X_0 \sim \text{Unif}(0, 1)$ (Malaspinas, 2016; Hui and Burt, 2015), then under the WFD-HMM the full data likelihood can be expressed as (Hui and Burt, 2015)

$$L(N) = f(Y_t, Y_0 \mid N, X_0, X_t)$$

$$= f(Y_t \mid X_t) f(Y_0 \mid X_0) f(X_t \mid X_0, N) f(X_0 \mid N)$$

$$= f(Y_t \mid X_t) f(Y_0 \mid X_0) f(X_t \mid X_0, N) f(X_0)$$

$$\propto \int_0^1 f(Y_t \mid X_t) \left( \int_0^1 f(X_t \mid X_0, N) f(Y_0 \mid X_0) f(X_0) dX_0 \right) dX_t$$

$$\propto \int_0^1 f(Y_t \mid X_t) \left( \int_0^1 f(X_t \mid X_0, N) f(X_0) dX_0 \right) dX_t$$

(4.1)

To perform the maximum likelihood estimation Hui and Burt (2015) makes the following assumptions regarding the components of 4.1:

1. $Y_t \mid X_t$ is distributed binomially

2. $f(X_t \mid X_0, N)$ and $f(X_0 \mid Y_0)$ are both beta distributed and

3. $\int_0^1 f(X_t \mid X_0, N) f(X_0) dX_0$ can be approximated as a beta distribution.

By constructing the likelihood in this way, Hui and Burt (2015) impose a beta-binomial distribution on the likelihood of $N$. With approximation by a known parametric distribution, the complexity of the numerical optimization of the maximum likelihood estimate is greatly reduced. Therefore, as a result the process
shows improvements in speed over more computationally intensive estimation methods of $N$ (Williamson and Slatkin, 1999; Anderson et al., 2000; Bollback et al., 2008). However, the assumption of a beta distributed DAF, $f(X_t \mid X_0, N)$, suffers from the limitation of not accounting for positive probability at the boundary. Thus, an extension of this maximum likelihood approach for estimating $N$ is to substitute the beta distributed DAF or approximate the entire interior integration problem of $\int_0^1 f(X_t \mid X_0, N)f(X_0 \mid Y_0)dX_0$ in equation 4.1 with the BWS or BWS-MLR distributions from section 4.1.3.

### 4.2 Discussion

Using the PRRSV next-generation sequencing studies as a motivating force for this investigation, the methods presented in this dissertation attempt to address the issue of utilizing time-serial genomic data to detect natural selection. First, we leveraged the PRRSV data across multiple pigs, genotypes and time-points, where all levels of genetic population structure were tested to identify whether the genetic variation was associated to the stratifications of the PRRSV data structure. The AMOVA demonstrated that the experimental trial effect is large, however, all covariance components were concluded to be significant. The importance of time in the assessment of the existence of genetic variation was further supported by the identification of temporal associations in genetic variation profiles across a collection of pigs from the PRRSV host study using the CMH test. Despite evidence of serial trends and accounting for population structure, one cannot make definitive assessments of the presence or magnitude of selective force because it cannot be differentiated from random drift.

Proper estimation of the effective population size has major implications on the utility of selection detection methods, however neither the AMOVA nor CMH tests in the PRRSV studies take this effect into account when testing for association. For instance, in cases where $N$ is large, allele frequency changes become less random and more deterministic, resulting in a lack of information to estimate $N$ altogether. Often, the effective population size and selection coefficient cannot be jointly estimated, causing complications for a large range of techniques that rely on prior information regarding the magnitude of $N$.

Secondly, in chapter 3 we explored testing methods to circumvent the need to explicitly estimate $N$ and the elapsed time $T$, namely the FITR. By applying normalizing transformations to a simulation scheme of highly skewed data representative of the PRRSV genomic data utilized in chapter 2, we overcame the violations of normality. In comparison to the untransformed data, the post-transformation results reduced skewness and lowered the excess kurtosis, as well as conserving the type-1 error rate, in the majority of
all the simulations evaluated. Situations where the original data did not exhibit any major violations of normality prior to transformation roughly remained unchanged post-transformation. Ultimately, we extend the utility of the FITR far beyond combinations of large \( N \), short \( T \) and intermediate initial allele frequency recommended by the original author (Nishino, 2013). However innate characteristics of the untransformed allele frequency distribution, particularly the sizable probability mass at the boundaries, limited the post-transformed distribution to only test one-tail alternatives to the null hypothesis of \( s = 0 \). We found that the choice of the normalizing transformations were highly dependent on the value of \( N \) and \( T \) in our extensive simulation study despite the FITR not explicitly depending on \( N \) and \( T \).

The inference of distinguishing natural selection from genetic drift within quickly evolving viral genomes remains a substantial challenge. Our methods rely heavily on the extension of the Gaussian approximation to the WF-HMM and exploiting this distributional form allows for use of well-developed modeling/estimation techniques. Notably, the stringent assumptions of both the evolutionary demographic scenarios and distributional form of the model will not apply to the vast majority of real-world applications. Despite its simplicity however, the characteristics of this framework provide a gateway for more statistically rigorous maximum likelihood models and hypothesis tests to naturally be extended from. In particular, one could imagine the development of a new FITR test that incorporates a joint distribution of reference loci to increase the power to pinpoint selected loci in the presence of linkage disequilibrium (§4.1.2). Although estimating \( N \) is not of direct interest in this research, the FITR’s ability to be adapted to account for fluctuations in population size is of primary value since the efficacy of the selective force is affected by population size. Lastly in recent years, methods have explored the application of alternative approximations to the distribution of allele frequency defined by the WF-HMM (Gautier and Vitalis, 2013; Sirén et al., 2011; Tataru et al., 2015; Hui and Burt, 2015). Since preliminary exploration tends to support that the beta approximations may be more suited to accommodate the various profiles that true allele frequencies can exhibit, it raises a question about how tests based on the normal frequency increments like FITR would fare against a comparable selection detection test constructed under the beta approximation. While we do not explore the development of a likelihood based neutrality test for beta increments, we offer an enhancement to the beta with spikes distribution (Tataru et al., 2016) to better estimate the spikes (point masses) as functions of the initial allele frequencies (§4.1.3).

This research presents more powerful methods to detect sites under the influence of selection, while identifying key areas for additional research. With continuous advances in high-throughput sequencing technologies, large scale genetic analysis allows for a better understanding of natural selection within evolving
populations. In regards to PRRSV, more insight into the role selection plays in virus survival can aid in identifying weakly selected loci that are correlated to neutral or strongly selected loci, recognizing recessive alleles and inferring the action of natural selection on gene regulation. More broadly, these findings can have direct implications to the ongoing development of pathogen resistance treatments and other widespread control protocols necessary to reduce the prevalence of PRRSV outbreaks around the world.
Figure 4.1: Difference between Harrell-Davis quantile estimates and standard normal quantiles for (a) lower tail quantiles and (b) upper tail quantiles across various normalizing transformations where $N = 1000$. Each row corresponds to a specific quantile and each column corresponds to a specific time lapse $T$. Estimates of the upper tail quantiles (b) tend to be more variable and underestimate the true standard normal quantiles than the lower tail quantile estimates.
Figure 4.2: The Hellinger distance between the WF-HMM DAF under (a) $N = 200$ and (b) $N = 1000$ and various approximations on a log scale. Approximations are truncated normal, beta with spikes, beta with spikes from Simulation A (beta with Hellams spikes) [see Table 3.2], and Box Cox transformation [see Table 3.1]. Continuous approximations were discretized into $N + 1$ equal-sized bins within the closed unit interval. In the case of the beta with spikes and beta with Hellams spikes models, the point masses $\hat{\gamma}_{A,0}$ and $\hat{\gamma}_{B,0}$ were added to the first and last bin, respectively. Darker blue hues indicate more distributional similarity to the WF-HMM DAF (Hellinger distance closer to 0) and darker red hues indicate less distributional similarity.


