2016

Studying Within-Host Viral Evolution Using Pooled Next-Generation Sequencing Data

Chase W. Nelson
University of South Carolina

Studying Within-Host Viral Evolution Using Pooled Next-Generation Sequencing Data

by

Chase W. Nelson

Bachelor of Arts
Oberlin College, 2010

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Biological Sciences

College of Arts and Sciences

University of South Carolina

2016

Accepted by:

Austin L. Hughes (deceased), Major Professor

Joseph M. Quattro, Major Professor

Robert Friedman, Committee Member

Ward Watt, Committee Member

Roger Sawyer, Committee Member

John A. Kupfer, Committee Member

Meredith Yeager, Committee Member

Lacy Ford, Senior Vice Provost and Dean of Graduate Studies

DEDICATION

To Austin L. Hughes, eternal mentor and friend. *Dwi'n dy golli di*.

To Eileen Snow Hicks, Evelyn Snow Kraak, and Elizabeth (Betty) Nelson, my grandmothers and teachers in the ways of the heart and mind. I will never, ever forget you.

To Lisa Kouchnerkavich and James Nelson, my parents, for their love; to Ben Kouchnerkavich, Jen Zeerip, and Missi Bastiaanse, my brother and sisters, for their camaraderie; and to Kevin Kouchnerkavich and the rest of my family for a lifetime of support.

To Jennifer Fill, Zachary Hoyer-Leitzel, Michelle Calkins, Elizabeth Ogle, Deb Tindall, Sara-Alicia Gonzalez, Vicky Saye Henderson, Terrance Henderson, Mindy Rawlinson, Cindy Vallone, Nicholas Loren, Anton Dela Cruz, Mark Altman, and innumerable other dear friends for their laughter, wisdom, inspiration, and affection on the journey.

ACKNOWLEDGEMENTS

I am most indebted in this work to Austin L. Hughes, for giving me almost unfathomable freedom, for sharing ideas and expertise with a selfless passion, and for responding to every question with alacrity; and to Wen-Hsiung Li for two summers of mentoring in Taiwan. I am also thankful for the generous help extended by Meredith Yeager, Wen-Hsiung Li, David O'Connor, Margaret Cirtain, Johannes Stratmann, Dan Graur, and many others following the passing of Professor Hughes. Finally, I am grateful for insightful discussions with April Hall, Joe Quattro, Meredith Yeager, Saravanan Rajabojan, Shrujan Amin, James Hussey, and (of course) my excellent Committee at the University of South Carolina.

This work was supported financially by National Science Foundation Graduate Research Fellowship DGE-0929297, a University of South Carolina Presidential Fellowship, and the University of South Carolina Department of Biological Sciences through two years of teaching assistantships and the Kathryn Hinnant-Johnson, M.D. Memorial Fellowship.

ABSTRACT

Pooled next-generation sequencing allows multiple genomes to be sequenced at once in a single sample, with the resultant single nucleotide polymorphism data giving reliable estimates of allele frequencies and population genetic parameters in a cost-effective manner. This approach has potentiated new opportunities for understanding the evolution of virus populations within individual hosts over the course of infection, where the sequencing of individual genomes is exceedingly difficult and impractical. However, evolutionary tools for analyzing the latest forms of pooled-sequencing data have been lacking. In this thesis, I first review next-generation sequencing and relevant molecular evolution topics, including the unique features of RNA viruses. I conclude that viruses, given their extremely fast replication rates and within-host population sizes, are ideal models for studying evolution by natural selection. Next, simple methods are devised for estimating nonsynonymous and synonymous nucleotide diversity from pooled next-generation sequencing data, without the need for inferring linkage. I introduce SNPGenie, a new bioinformatics tool for applying these methods to any pooled or individual variant data. Finally, I use SNPGenie to address topics of both practical and theoretical interest in the evolution of simian hemorrhagic fever viruses (*Arteriviridae*) infecting red colobus monkeys (*Procolobus rufomitratus tephrosceles*), including fundamental questions regarding the effective population sizes of, the mutation rates experienced by, and the modes and efficacy of natural selection acting on within-host viral populations.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

D Tajima's statistic for detecting selection. If positive, a paucity of rare alleles is observed, implying positive selection; if negative, an excess of rare alleles is observed, implying negative (purifying) selection.

H Gene diversity. The probability that two alleles chosen at random from a population differ. This is equivalent to expected heterozygosity in a randomly mating diploid population.

N Actual population size. Because real populations do not adhere to assumptions such as constant population size, population genetics theory must account for random effects by instead defining a population's effective population size (N_e).

N_e Effective population size. This is the size of the ideal population that would behave in an evolutionary sense like a given population of actual size N, after accounting for random effects such as fluctuations in population size and spatial structuring. In almost all instances, $N_e << N$.

Π The mean number of pairwise differences in a population of sequences, uncorrected for sequence length.

π Nucleotide diversity. The mean number of pairwise differences per site in a population of sequences. This can be calculated separately for nonsynonymous and synonymous sites, yielding π_N and π_S, respectively.

S The number of segregating (polymorphic) sites in a population of sequences.

θ Population genetic parameter equal to $2N_e\nu$ for haploid populations and $4N_e\nu$ for diploid populations.

θ_π Estimate of θ based on the mean number of pairwise differences within a population of sequences, Π; this is Tajima's estimate. Used in calculating Tajima's D.

θ_S Estimate of θ based on the number of segregating sites, S; this is Watterson's estimate. Used in calculating Tajima's D.

ν Mutation rate per site. Whether this is per generation or per unit time (*e.g.*, year) is defined situationally in the text.

LIST OF ABBREVIATIONS

kb ..... kilobase

NGS ..... next-generation sequencing

NOL ..... non-overlapping

nonsyn. ..... nonsynonymous

nt ..... nucleotide

OL ..... overlapping

RC ..... red colobus (*Procolobus rufomitratus tephrosceles*)

SHFV ..... simian hemorrhagic fever virus

SNP ..... single-nucleotide polymorphism

syn. ..... synonymous

CHAPTER 1

BACKGROUND

1.1 Pooled Next-Generation Sequencing

Over the past decade, a host of new DNA sequencing tools, dubbed "next-generation" sequencing (NGS) technologies, have for the first time allowed millions to billions of nucleic acid sequences to be processed in parallel directly from sequence fragment libraries (Metzker 2010). By tremendously increasing the speed and accuracy, while concurrently decreasing the cost of sequencing, NGS platforms such as Illumina and Roche/454 have potentiated unprecedented sequencing and resequencing efforts (Mardis 2008; Shendure and Ji 2008). These include initiatives both to characterize genetic variation within a species (Nielsen et al. 2011) and to quantify gene expression differences through transcriptomic analyses using RNA-seq (Martin and Wang 2011). Massive-scale examples of the former, involving sequencing of large numbers of individual genomes, include the 1000 Genomes Project for humans (Auton et al. 2015; Sudmant et al. 2015) and the 1001 Genomes Project for *Arabidopsis* (Cao et al. 2011).

More recently, it has been realized that population genetic characterization need not require the individual sequencing of multiple genomes, but can instead be accomplished through the pooling of those genomes from separate individuals into a single representative sample. Analyses of natural isolates and samples constructed with known frequencies of single nucleotide polymorphisms (SNPs) have established that this "pooled-sequencing" approach gives accurate estimates of allele frequencies above at least the 1% level, for both Roche/454 (Ingman and Gyllensten 2009; Becker et al. 2012; Rellstab et al. 2013) and Illumina technologies (Wright et al. 2011; Dudley et al. 2014), and statistical developments can further improve accuracy (Futschik and Schlötterer 2010; Lynch et al. 2014).

Illumina sequencing, which is the technology used in my studies, works by utilizing the so-called "sequencing by synthesis" approach (for overviews, see Mardis 2008 and Shendure and Ji 2008). First, libraries are constructed by fragmenting the source DNA and simultaneously ligating adaptor sequences to the fragment termini. In the case of RNA genomes, this step is preceded by reverse transcription, yielding complementary DNA (cDNA) that can be used for library construction. Following adaptor ligation, additional sequence motifs are added to allow sample identification and binding of fragments to the Illumina flow cell. Flow cells refer to lanes on a sequencing plate, which are seeded with oligonucleotides that are complementary to the newly ligated fragment ends. The library is washed over the flow cell in sufficiently low concentrations so as to allow binding of separate fragments in distinct locations on the plate, nevertheless allowing millions of fragments to bind. The plate- and fragment-bound oligonucleotide sequences are then extended by DNA polymerase, resulting in millions of plate-bound sequence fragments, and further amplified by bridge PCR to form clusters of identical sequences on the plate (roughly 1,000 copies per cluster). These clusters are then sequenced by repeated addition of a mix of all four of the DNA nucleotide bases: adenine (A), cytosine (C), guanine (G), and thymine (T). Each base is chemically linked to its own fluorescent label and a 3'-OH group, which prevents more than one base from being incorporated at a time. The plate is imaged to record the color of the base added at each distinct plate location (cluster), and then washed with an agent that removes the fluorescent label and 3'-OH group, allowing further addition and sequencing in subsequent cycles. The primary limitation of the technology is the short length of sequence fragments, which results from incomplete cleavage of fluorescent

labels and/or 3'-OH groups; as a result, the quality of base determination (*i.e.*, base "calling") diminishes further into the sequenced fragment, otherwise known as a read (Shendure and Ji 2008). Fortunately, this can be substantially overcome by paired-end sequencing, and both Illumina's Hi-Seq and Mi-Seq platforms presently boast high-quality paired-end reads exceeding 150 nt. Error rates of ~1% are nevertheless common, mostly single nucleotide substitutions (rarely insertions or deletions), necessitating the use of base quality scores and sophisticated algorithms to determine legitimate variant calls (Nielsen et al. 2011).

Whatever the NGS technology, the establishment of pooled-sequencing as a reliable approach to population genetic research presents a unique opportunity for studying the evolution of within-host viral populations. By circumventing the painstaking process of molecular cloning and sequencing of individual viral genotypes, pooling instead allows viral genomes to be isolated from infected tissues and amplified for sequencing *en masse* with no loss of accuracy, and an arguable gain in population-level resolution (Wright et al. 2011). SNP calling approaches can be used with sufficiently stringent criteria (*i.e.*, base quality requirements and minimum allele frequency cutoffs) to allow the most important genetic variation to be identified with confidence. Moreover, important population genetic parameters such as nucleotide diversity (see Section 1.1 and Chapter 2) do not depend on very low-frequency SNPs, so their utility is not limited even by quite stringent base calling criteria (as in Chapter 4). In cases where the inoculum (infecting) sequence is known, pooled-sequencing can determine how a virus has evolved at distinct sites in a host. Where longitudinal data are available, comparisons of a within-host viral population between two or more distinct time points can reveal viral dynamics

over time and the likely selective pressures at play. Finally, even when the inoculum is not known, analyses of natural isolates can be used to characterize viral polymorphism in ways that give evolutionary insights as to the relative roles of natural selection and genetic drift.

1.2 Molecular Population Genetics

Evolution involves change in the genetic makeup of populations over time. DNA—or RNA, in the case of many viruses—houses this genetic information, stored as a linear sequence of the four nucleotide bases (RNA genomes use uracil, or U, in place of T). A population's genome refers simply to its consensus or most common DNA (or RNA) sequence. Each individual within a diploid population (*e.g.*, most mammals) stores two unique copies of the genome in each cell, while haploid individuals (*e.g.*, viruses) contain only one copy. Genomes themselves are organized into regions called *genes*, each of which encodes a specific molecule that plays some functional role for the organism. The most-studied gene product is the protein, which in the case of DNA genomes is formed first by transcription of an RNA copy of the gene, then by translation of the RNA into protein. While it has traditionally been thought that gene function relies chiefly on the expression of proteins, large-scale research efforts such as the ENCODE (ENCyclopedia Of DNA Elements) Project have raised awareness of a new world of functional RNA molecules that we are only beginning to understand (Birney et al. 2007; Gerstein et al. 2007; Myers et al. 2011). It is too early to make a judgment about the fraction of mammalian genomes that encode functional genes, and the size of the subset of those genes which encode functional RNAs; however, interspecies conservation suggests that

the total cannot be much more than 5-10% in humans (Lindblad-Toh et al. 2011; Graur et al. 2013).

In diploids, genes are typically separated by much longer intergenic regions which may play structural or regulatory roles but are thought to lack other functionality in most instances. On the other hand, haploid genomes are relatively compact, with the genomes of viruses in particular housing few if any intergenic regions. RNA viruses are especially compact, their genes often overlapping and their nongenic regions being limited to small segments at the genomic termini.

Each copy of the genome within a population typically differs from every other copy as a result of mutations, random errors introduced either during genome replication or else by chemical or radioactive perturbation. When a gene incurs a mutation, a variant form called an *allele* results. An individual's unique genome sequence (*i.e.*, its collection of alleles) is referred to as its *genotype*, while its *phenotype* refers to its expressed characteristics, physical traits, and physiological qualities. Importantly, while mutations by definition change an organism's genotype, they may or may not have an effect on the phenotype. Those that affect the phenotype, in turn, may or may not affect *fitness*, defined here as the organism's potential for long-term (viable) reproductive success over the course of its life cycle. For viruses, this is often simply estimated as a replication rate. With respect to fitness, mutations may be characterized as deleterious (decrease fitness), neutral (no effect), or beneficial (increase fitness). While mutations can have fitness effects ranging from lethal to necessary for survival, the majority are slightly deleterious (Eyre-Walker and Keightley 2007).

Mutational fitness effects allow the frequencies of mutations to be influenced by the action of natural selection. Deleterious mutations by definition cause their carriers to leave fewer offspring, on average, as compared to individuals with the mutation-free genotype. The result is a phenomenon called *negative* (*purifying*) *selection*, in which the frequency of a mutation decreases in a deterministic fashion. Neutral mutations, on the other hand, are subject to *genetic drift*, increasing or decreasing in frequency at random. These can also hitchhike to either extinction or fixation with other mutations which are themselves under selection. Finally, beneficial mutations by definition cause their carriers to leave more offspring, on average, as compared to individuals with the mutation-free genotype. The result is *positive* (*Darwinian*) *selection*, in which the frequency of a mutation increases in a deterministic fashion. Over a number of generations, a beneficial mutation may increase to a frequency of 100%, *i.e.*, every individual in the population carries the mutation in its genome. The mutation is then said to have reached *fixation* through an evolutionary substitution—by which is meant the substitution of one (pre-mutation) allele by another (beneficial mutant) allele in the population. The speed at which a mutation is eliminated by purifying selection, or promoted by positive selection, is dependent on the magnitude of the mutation's fitness effect.

It has long been recognized that, owing largely to the small (finite) sizes of real populations in the natural world, deterministic processes such as natural selection do not always play a leading role in driving evolutionary substitutions (Wright 1931; Li 1997; Lynch 2007a; Lynch 2007b; Hughes 2008; Koonin 2009). Moreover, some have suggested that there is a rate limit, probably imposed by a species' finite reproductive capacity, on the speed at which selection-driven substitutions can occur (*e.g.*, see Nelson

2015). Originally proposed by Haldane (1957, 1960) as the "cost of selection," this idea has attracted considerable criticism since its proposal (*e.g.*, see Brues 1964; Ewens 1979; Wallace 1988; Wallace 1991; Woodruff et al. 1996; Watt 2003; Woodruff et al. 2004). Whatever the true rate limit may be, these ideas produced a perceived need for another mechanism of evolution, not deterred by the limitations of natural selection, which can explain the numbers of genetic difference observed between species, as well as the amount of polymorphism currently observed within populations. The result was the *neutral theory of evolution*.

The neutral theory hoped to explain the large numbers of genetic differences observed between species by relying on relatively ubiquitous neutral mutations, rather than on the relatively rare beneficial mutations, as previous approaches had done. The neutral theory's central claims are that the majority of evolutionary substitutions occur by random genetic drift, and that the majority of polymorphisms are selectively neutral alleles (Kimura 1968; Kimura 1979; Kimura 1983; Ohta 1992; Hughes 2008; Nei et al. 2010). In other words, polymorphism is a transient phase of the mechanism by which most molecular evolution takes place (Hughes 1999).

The hypotheses represented by the neutral theory have been the focus of much heated debate among so-called "neutralists" and "selectionists" working in evolutionary biology (Dietrich 1994). Rather than engage that debate here, I instead briefly discuss those aspects of the neutral theory that may be relevant in important ways to viral evolution. First, the neutral theory recognizes that an inverse relationship exists between the effective size of a population, N_e, and the range of mutations dominated by genetic drift, called nearly-neutral mutations. As compared to N, which represents the actual

(census) population size, N_e is the size of the ideal population (usually much smaller than N) which behaves in an evolutionary sense like a given population of size N, *after accounting for* random effects such as fluctuations in population size, unequal contributions of different individuals to the next generation, and spatial structuring. More specifically, following Wright (1931), the neutral theory states a mutation will be nearly neutral if its fitness effect falls approximately within the range $-1/N_e$ to $1/N_e$ (Kimura 1983). Thus, the larger the population, the narrower this window of nearly neutral fitness effects, and the higher the number that will be dominated by selection in the evolutionary process. This can be especially important for viruses during the process of transmission to new hosts, which can be accomplished by very few viral particles, or during the infection of new micro-environments within a single host. Thus, the neutral theory emphasizes demographic events such as population bottlenecks (*i.e.*, extreme reductions in population size) in the course of evolution. To the extent that the elimination of individuals during such an episode is random with respect to fitness, genetic drift can accomplish a great deal of work, such as the fixation of rare alleles. Finally, the neutral theory makes several predictions about genetic variation within a population, including: (1) purifying selection against deleterious polymorphism should be common; (2) neutral polymorphism should increase over time; and (3) larger populations should harbor more polymorphism (see Chapter 2).

Prediction (1) above deserves further comment. Within protein-coding genes, nucleotides are arranged for functional purposes in triplets, called codons, each of which encodes a single amino acid to be incorporated into a protein product. For example, the codon CCC encodes the amino acid proline. The sites within a codon can be further

categorized as either nonsynonymous or synonymous. In human language, two words which are *not* synonymous have different "meanings"; in the same way, two nonsynonymous nucleotides in a genome encode different amino acids. For example, replacing the first C in CCC with A, G, or T results in a different amino acid (ACC encodes threonine; GCC encodes alanine; TCC encodes serine). For this reason, the first site of CCC is nonsynonymous. On the other hand, some codon sites are synonymous, because changes in the nucleotide do not change the amino acid product. For example, replacing the third C in CCC with A, G, or T does not change the amino acid (CCA, CCC, CCG, and CCT all encode proline). Some sites are only fractionally nonsynonymous, in the sense that only some of the possible nucleotide replacements at the site will change the amino acid. Various methods exist for dealing with such sites, ranging from very simple approaches to ones which involve parameters such as the ratio of transitions (mutations between A and G, the purines, or between C and T, the pyrimidines) to transversions (all other mutations) (Nei and Kumar 2000).

Nonsynonymous mutations are able to affect protein structure directly by changing the amino acid encoded, and are thus subject to natural selection acting on the phenotype. By comparison, synonymous mutations are generally "silent," although some notable exceptions exist (*e.g.*, see Komar 2007; Hunt et al. 2009; Kimchi-Sarfaty et al. 2016). The prediction that purifying selection is common can therefore be tested by comparing polymorphism at nonsynonymous and synonymous sites, since purifying selection (if common) should act to decrease nonsynonymous polymorphism as opposed to synonymous polymorphism, the latter of which accumulates relatively neutrally. However, approximately 75% of all sites in a random DNA sequence are

nonsynonymous; thus a simple comparison of the number of polymorphisms of each type is not a fair one, since nonsynonymous mutations are three times as likely as synonymous ones. To correct for this, the number of nonsynonymous mutational differences among a pair of sequences, m_N, is normalized by the number of nonsynonymous sites, n_N. Likewise, the number of synonymous differences among a pair sequences, m_S, is normalized by the number of synonymous sites, n_S. For a population of N haploid sequences, there are ${}_N C_2 = (N^2 - N)/2$ pairwise comparisons among the genome copies. The number of normalized nonsynonymous differences for each pairwise comparison constitutes an estimate of nonsynonymous polymorphism:

$$d_N = \frac{m_N}{n_N} \qquad \textbf{[equation 1.1]}$$

Likewise, the number of normalized synonymous differences constitutes an estimate of synonymous polymorphism:

$$d_S = \frac{m_S}{n_S} \qquad \textbf{[equation 1.2]}$$

The mean of all normalized differences for all $(N^2 - N)/2$ pairwise comparisons is then equivalent to the population genetic parameter nucleotide diversity, denoted π (Nei and Li 1979). More specifically, mean within-population d_N is equivalent to nonsynonymous nucleotide diversity π_N, and mean within-population d_S is equivalent to nonsynonymous nucleotide diversity π_S. Estimates of these parameters will be used extensively in subsequent chapters.

Given estimates of π_N and π_S for a population, it becomes possible to test for purifying selection by comparing their values. Because selection affects nonsynonymous but generally not synonymous sites, when purifying selection is widespread, we expect $\pi_N < \pi_S$; when genetic drift dominates, we expect $\pi_N = \pi_S$; and when positive selection is widespread, we expect $\pi_N > \pi_S$. As it turns out, most genes in most species exhibit the pattern $\pi_N < \pi_S$, vindicating this prediction of the neutral theory (Hughes 1999; Hughes 2007). This holds equally for all viral genomes (Chapter 2; Holmes 2009).

The fact of widespread purifying selection is what enables perhaps the most important tool of evolutionary bioinformatics: sequence comparison. More specifically, if most functional genetic sequences are under purifying selection, then they experience evolutionary constraint. Such constraint preserves the sequences of functionally important regions over time as compared to sequences lacking function, allowing them to be aligned according to their similarity. Thus, degree of similarity can be used as an indicator of function, and can even help researchers to infer the function of newly discovered genes from their sequences alone using tools such as BLAST (Altschul et al. 1990). If functionally important genes were instead constantly under positive selection over the course of evolution, they would change very rapidly, and the insights afforded by sequence comparison would not always be possible (Hughes 2011).

Although purifying selection is the norm, important instances of positive selection do exist. When selection favors multiple repeated changes in a genomic region, the pattern $\pi_N > \pi_S$ can result. Perhaps the most important instance of this—and one which is highly relevant to viral evolution—is the case of the vertebrate major histocompatibility complex genes, to which we now turn.

1.3 The Major Histocompatibility Complex

The surfaces of all nucleated cells in vertebrates are studded with major histocompatibility complex (MHC) class I receptors. The genes encoding these receptors, known as human leukocyte antigen (HLA) genes (Lawlor et al. 1990), have long been known as the most polymorphic loci observed in animals (Klein and Figueroa 1986). It has also been known that these receptors play a key role in the immune response by displaying peptide fragments (small pieces of proteins) to the immune system's cytotoxic (CD8+) T cells (Klein 1986). This is accomplished through a random sampling of the proteins present inside a cell, which are sliced into fragments within LMP+ proteasomes, transported to the endoplasmic reticulum, attached to MHC class I receptors, and shuttled to the cell surface for display (Hughes 1999). If circulating cytotoxic T cells encounter a cell displaying an MHC class I molecule that is complexed with a non-self peptide fragment, the T cell can then set in motion the infected cell's destruction.

The reason for high levels of MHC polymorphism was not always clear, and early hypotheses were numerous, including the idea that the MHC gene loci were mutational hotspots (Klein 1978). Hughes and Nei (1988) were the first to conclusively establish positive selection as the culprit. Following a handful for fortuitous developments, including the demonstration that different MHC alleles are able to bind different peptide fragments (Zinkernagel and Doherty 1974) and the determination of the MHC receptor's molecular structure (Bjorkman et al. 1987), Hughes and Nei (1988) were able to validate a hypothesis first proposed by Doherty and Zinkernagel (1975). This hypothesis invokes a role for the overdominant type of positive selection, otherwise known as heterozygote advantage, in maintaining the polymorphism of the MHC locus. Rather than driving

evolutionary substitutions, overdominant selection constitutes a form of balancing selection in which diversity, rather than fixation, is favored. One case, sickle cell anemia, had already been well established. Here, a mutant allele in its homozygous state (two identical alleles in a diploid organism) causes the debilitating disease sickle cell anemia, but in its heterozygous state (two differing alleles in a diploid organism) provides protection against the malarial pathogen. As a result, the mutant allele is maintained in regions where malaria is prevalent, despite the adverse effects of the allele in homozygotes.

Hughes and Nei (1988) used comparisons among 12 available human and 8 available mouse DNA sequences to establish the operation of the same evolutionary mechanism at the MHC loci. For both human and mouse, π_N was significantly greater than π_S in intralocus comparisons. Moreover, π_S was no greater at the MHC loci than other parts of the genome, ruling out the alternative hypothesis of an elevated mutation rate. Subsequent work established that the pattern of $\pi_N > \pi_S$ occurs precisely at the MHC residues responsible for binding the pathogen-derived peptide fragment, *i.e.*, the MHC peptide binding region, and that nonsynonymous changes which alter amino acid electric charge explain most of this pattern (Hughes et al. 1990). Hughes and Yeager (1998) drew upon accumulating sequence data from different species, and several alternative lines of evidence including trans-species polymorphism, to further support this case. Overdominant selection thus makes an elegant explanation for the polymorphism of the MHC loci—and one which takes their function into account (Hughes 1999).

Pathogens must co-evolve with the immune systems of their hosts (Howard 1991). Each exerts strong selective pressures on the other, as immune surveillance

competes with immune escape. Thus, the diversity observed at an individual's MHC loci should be reflected in the population-level diversity of an infecting pathogen. If particular pathogenic proteins allow MHC binding and presentation, the infectious agent may do well if the peptide mutates beyond the recognition of the immune system, while simultaneously allowing the pathogen to remain viable. In this thesis, I take the point of view of the infecting pathogen, and we shall turn our attention specifically to viruses.

1.4 Viruses

Viruses are intracellular *pathogens* (disease-causing agents) whose genomes consist of either single- or double-stranded DNA or RNA. My focus will be the Arteriviruses (*Arteriviridae*), single-stranded (non-segmented) positive-sense RNA viruses infecting mammals. Because host cells do not express RNA-dependent RNA polymerases, RNA viruses carry genetic information for their own polymerase in order to replicate. Positive-sense RNA genomes read as a messenger RNA (mRNA) molecule, and are thus ready for immediate translation upon infection. However, in Arteriviruses, only translation of the first two open reading frames (ORFs) occurs upon cell entry, the protein products of which are cleaved and assembled into a replication and transcription complex (RTC). The RTC then engages in negative-sense strand synthesis (beginning at the 3' end of the genomic RNA), producing both full-length and subgenomic-length copies of the RNA genome, the latter containing only some fraction of the genome starting from the 3' end. The full-length copies are again transcribed into positive-sense genomic copies to be incorporated into viral progeny, whereas the subgenomic-length copies are transcribed

into mRNA molecules for use in expression of the remaining ORFs (Snijder et al. 2013). For more details on the Arterivirus genome, see Chapter 4.

Considerable debate has taken place over whether viruses should be classified as living organisms (Raoult and Forterre 2008). The reason for this is quite simply that viruses are obligate parasites—they are not free living, and require a host cell for completion of their life cycle. However, this is not unlike obligate intracellular bacteria (Brüssow 2009), which are considered living, and indeed all living things are dependent on some sort of environment for surviving and obtaining energy. For our purposes, I bypass this question and note only that viruses possess all the characteristics necessary to study biological evolution: replication, inheritance, and heritable genetic variation that arises as a result of mutation.

In fact, several characteristics make RNA viruses especially amenable to evolutionary study (for an overview, see Holmes 2009). First, they have extremely high mutation rates, ranging from about 10^{-6} to 10^{-4} mutations per nucleotide per cell infection, with point mutations being about 4 times more common than insertions and deletions (Strauss and Strauss 2008; Sanjuán et al. 2010). This is the result of the RNA-dependent RNA polymerase, which lacks the proofreading capabilities of DNA polymerase. Since some viruses may undergo several rounds of copying within a single cell, the actual rate per replication may be somewhat lower. Importantly, although estimates of the mean mutation rate exist, the *distribution* of the mutation rate is not known. Determining this distribution will be crucial to understanding the evolutionary trajectories taken in RNA virus evolution. For example, a symmetric and peaked distribution would imply the production of numerous viral particles with the mean number of mutations, while an

alternative (*e.g.*, bimodal) distribution might imply a disproportionately large number of viral progeny having either very few or very many mutations, the latter potentially enabling fitness valleys to be traversed but simultaneously imposing a heavy mutational burden.

The distribution of mutational fitness effects in viruses is bimodal, with most mutations being either lethal or nearly-neutral (Eyre-Walker and Keightley 2007). One study of single-stranded RNA and DNA viruses estimated that 20-41% of all mutations are lethal; of the remaining (viable) mutations, the mean fitness effect ranged from -0.103 to -0.132 (Sanjuán 2010). Thus, the extremely high mutation rate of RNA viruses exerts a tremendous burden of deleterious mutations. This fact has informed certain therapeutic strategies, which take the approach of elevating a virus' mutation rate (*e.g.*, through application of the drug ribavirin) to induce lethal mutagenesis (Bull et al. 2007). In similar fashion for the host, mutation accumulation may help to explain the gradual deterioration of immune cells during HIV progression (Galvani 2005).

RNA viruses are able to cope with the deleterious effects of their extremely high mutation rates mainly through their large population sizes and replication rates. In Arteriviruses infecting red colobus monkeys, viremia (blood viral load) has been measured to vary from 3.4×10^4 to 1.9×10^8 viral particles per mL (see Chapter 4). Although the number of virions produced through budding has not been measured for Arteriviruses to date, burst sizes (number of viral progeny per cell infection) estimates for other RNA viruses have centered on approximately 10^4, being approximately 10^4 for polioviruses (Kew et al. 2005) and 4.0×10^4 to 5.5×10^4 for simian immunodeficiency virus (Chen et al. 2007). Within-host viral population size thus tends to exceed the

number of hosts infected worldwide (Holmes 2009). Although viruses tend to experience a population bottleneck upon transmission (*i.e.*, infection of new hosts), their enormous replication rates can soon result in N_e values large enough to allow the action of selection, as for viruses which undergo persistent asymptomatic infections, like the Arteriviruses of African monkeys.

Another characteristic of RNA viruses that offsets their high mutation rates is their small genome sizes. One hypothesis states that the upper limit of genome size is approximately the reciprocal of the mutation rate, which for a mutation rate of 10^{-4} would be 10,000 nt (Eigen 1992). Roughly consistent with this, RNA virus genomes range in size from 2,500 to 31,500 nt, with a mean of approximately 10,000 nt (Holmes 2009). This is presumably because a genomic deleterious mutation rate of 1 per generation would be evolutionarily unsustainable, leading to mutational meltdown. Extra nonfunctional ("junk") genetic material might not be subject to high rates of deleterious mutation, but its accumulation may be limited by other factors, such as the energy burden and increased replication time associated with maintaining excess genetic material (Lynch 2007a). Arteriviruses are themselves among the largest RNA viruses, having genome sizes approaching 16,000 nt in length.

Holmes (2009) has described RNA viruses as constituting "some of the best-equipped laboratories to study evolution by natural selection." The reasons for this should now be clear. RNA viruses experience extremely high mutation rates, such that nearly every new virion produced obtains a new mutation. This sets them apart from even double-stranded DNA viruses, such as herpesviruses, which take advantage of DNA proofreading capabilities and evolve relatively slowly (Cullen et al. 2015). Possibly to

offset this, their burst sizes are on the order of their genome size, ensuring that at least some viral progeny will be free of lethal or deleterious mutations after every round of cell infection. Once infection is established, their within-host population sizes can range from 10^4 to 10^8 per mL of blood in the case of Arteriviruses, enabling extremely effective natural selection in cases of persistent infection. Finally, their small and streamlined genomes not only make them amenable to study, but also afford a very close genotype-to-phenotype relationship. The result is that the phenotype very accurately "advertises" the viral genotype, such that most viral traits have high heritability and thus respond well to selection.

1.5 Conclusion

Next-generation sequencing analyses involving pooled samples allow reliable population genetic data to be obtained from single sequencing runs. This approach is particularly advantageous for the study of viruses, where the sequencing of sufficiently numerous individual genomes from extremely large populations is costly and impractical. RNA viruses in particular exhibit very high mutation rates, large population sizes, and high replication rates, making them excellent study systems for evolution by natural selection. Unfortunately, statistical approaches and widely available tools for undertaking evolutionary studies with pooled-sequencing data have not been readily available until lately. Toward this end, I present a simple method, based on that of Nei and Gojobori, for estimating the population parameter nucleotide diversity (π) from pooled NGS data in Chapter 2. Chapter 3 goes further to instantiate and expand this method as part of the open-source software SNPGenie. Finally, Chapter 4 applies SNPGenie to study the

largest pooled-sequencing Arterivirus dataset to date, effectively characterizing the within-host evolutionary dynamics of the virus and developing a framework for evolutionary modeling, unsupervised epitope discovery, and mutation rate estimation. Future prospects are then briefly explored.

CHAPTER 2

WITHIN-HOST NUCLEOTIDE DIVERSITY OF VIRUS POPULATIONS: INSIGHTS FROM NEXT-GENERATION SEQUENCING[1]

[1] Nelson CW, Hughes AL. 2015. Within-host nucleotide diversity of virus populations: insights from next-generation sequencing. *Infection, Genetics and Evolution* 30:1-7. Reprinted here in modified form with permission of publisher; see Appendix A.

2.1 Introduction

Because of the rapid generation times and high mutation rates of most viruses, the virus population infecting an individual host can accumulate substantial genetic diversity over the course of infection. This diversity is in turn subject, like genetic diversity in any biological population, to the processes of natural selection and random genetic drift, which determine whether individual variants increase or decrease in frequency. Thus, the viral population infecting an individual host is subject to an evolutionary process. This evolutionary process may be important for the persistence of viral infection; for example, the host immune system may selectively favor viral variants that evade immune recognition. For this reason, understanding within-host viral evolution has been a major focus of research aiming to understand the mechanisms by which certain viruses, such as human immunodeficiency virus 1 (HIV-1) and hepatitis C virus (HCV), evade clearance by the host immune system and thus establish persistent infections.

In spite of the importance of understanding within-host evolution of virus populations, it has been difficult to study this process until recently. The advent of so-called "next-generation" sequencing (NGS) technologies, with their potential to survey thousands of viral sequences from a given host, has dramatically improved our ability to characterize within-host sequence diversity in viral infections. NGS has been applied to address such questions as overall viral diversity within-hosts (Wright et al. 2011; Lauck et al. 2012); evolution of T-cell epitopes under selection by the host immune system (Bimber et al. 2010; Hughes et al. 2010; Hughes et al. 2012; Mudd et al. 2012; O'Connor et al. 2012; Walsh et al. 2013); response of viruses to selection imposed by antiviral drugs (Wang et al. 2007; Cannon et al. 2008; Le et al. 2009; Hedskog et al. 2010);

differences between virus subpopulations infecting different host cell types (Rozera et al. 2009); and population bottlenecks in infection (Wang et al. 2010).

Here we discuss statistical methods for using NGS data to understand nucleotide sequence diversity of within-host viral populations, with particular emphasis on the comparison of nonsynonymous (amino acid-altering) and synonymous ("silent") nucleotide diversity in coding regions. NGS studies of within-host virus diversity use pooled samples, *i.e.*, the genetic material of multiple individuals pooled in a single sample, as opposed to sequencing individual viral genomes separately. Besides saving costs, the sequencing of sufficiently large pools has been shown to give more accurate estimates of population genetic parameters than those obtained from individual sequencing (Futschik and Schlötterer 2010). Such studies can be categorized as follows: (1) *targeted NGS*, using primers that amplify a specific short region of the viral genome, such as a specific T-cell epitope, thereby providing the complete sequences of haplotypes spanning that region (Bimber et al. 2010); or (2) *genome-wide NGS*, using sets of primers (*e.g.*, random hexamers) designed to obtain sequence information across all or most of the viral genome (Hughes et al. 2012; Wilker et al. 2013; Bailey et al. 2014). In the former type of study, standard methods of statistical analysis of sequence data (Nei and Kumar 2000) are directly applicable, including the estimation of nonsynonymous and synonymous nucleotide diversity and even phylogenetic tree reconstruction. However, because the sequence reads produced by NGS are short and thus provide limited information, phylogenetic trees are often poorly resolved in the case of targeted NGS.

In the case of genome-wide NGS, traditional techniques of sequence analysis are not directly applicable because of the lack of knowledge of haplotypes. Except when two

single nucleotide polymorphisms (SNPs) occur in the same short read, these methods do not provide any direct evidence regarding the phase of SNPs, *i.e.*, whether or not they occur together in the same haplotype. In some studies, determining haplotypes may be sufficiently important that researchers may want to make use of statistical methods for inferring haplotypes by assembling sequence reads (Beerenwinkel and Zagordi 2011). However, it is uncertain that haplotype inference will always be possible in the case of within-host viral populations, where all or most haplotypes may be very closely related and parallel mutations and recombination may obscure haplotype identities. Moreover, whenever haplotype inference is used, it must be kept in mind that any further inferences that rely on that inference remain conditional upon its accuracy.

For this reason, it may be useful in the case of whole-genome NGS to make use of methods that estimate population-level sequence parameters without the need to infer haplotypes. Here we discuss the theoretical basis of such methods and some examples of their application. We then briefly address the potential of these approaches for addressing some important theoretical and applied issues in the biology of viruses. As a specific example, we discuss how application of these approaches may provide data that will shed light on the relevance of the "quasispecies" model for understanding within-host evolution of viral populations.

2.2 Nucleotide Diversity

Nucleotide diversity (π) represents an important property of populations of nucleic acid sequences. In order to estimate nucleotide diversity in a population, we first take a random sample of n sequences from the population. Between each sequence and each

other sequence, we estimate d_{ij}, the number of nucleotide substitutions per site. A number of models are available for estimating d_{ij}, correcting for multiple hits and taking into account the effects of base composition bias and transitional bias (Nei and Kumar 2000). In the case of within-host virus populations, d_{ij} values are generally quite low (usually much less than 10%), and therefore the effect of these corrections will be very slight; thus, the uncorrected proportion of nucleotide differences between sequences often provides an adequate estimate of d_{ij}. Nucleotide diversity (π) is estimated by the mean d_{ij} for all $(n^2 - n)/2$ possible pairwise comparisons among sequences; *i.e.*,

$$\pi = \sum_{i<j} \frac{d_{ij}}{(n^2-n)/2} \qquad \textbf{[equation 2.1]}$$

In the case of coding sequences, important evolutionary information can be gained by estimating nucleotide diversity separately for nonsynonymous and synonymous sites. First, we estimate for each pair of sequences the number of nonsynonymous substitutions per synonymous site (d_N) and the number of synonymous substitutions per nonsynonymous site (d_S) (Chapter 1). In addition to correction for multiple hits, there exist a variety of methods for estimating d_N and d_S that also take into account nucleotide content and transitional bias (Nei and Kumar 2000). In the case of within-host viral populations, since the degree of sequence divergence is usually slight, the use of complicated models for estimating d_N and d_S has little effect on the results. Thus, a simple method, such as that of Nei and Gojobori (1986), usually provides adequate results. Note that complex methods for estimating d_N and d_S, such as likelihood methods, generally estimate nucleotide frequencies and other such parameters from the sequences

themselves; thus, this procedure can be positively misleading when the sequences analyzed are short, because the stochastic error of these estimates will be very high in the case of short sequences. These complex methods for estimating d_N and d_S should therefore be avoided in the analysis of short sequences, as in targeted NGS, or in the estimation of d_N and d_S in sliding windows along a gene.

In a population of sequences, let d_{Nij} be the estimate of d_N between sequences i and j. The nonsynonymous nucleotide diversity (π_N) is estimated by substituting d_{Nij} for d_{ij} in equation 2.1. Similarly, let d_{Sij} be the estimate of d_S between sequences i and j. The synonymous nucleotide diversity (π_S) is estimated by substituting d_{Sij} for d_{ij} in equation 2.1.

Selectively neutral nucleotide diversity provides an estimate of the population parameter θ, which is proportional to the product of the effective population size (N_e) and the mutation rate (v) per generation (Li 1997; Nei and Kumar 2000). This relationship holds under the assumptions of the infinite-sites model of population genetics, when mutation and drift are in equilibrium (Nei and Kumar 2000). Since synonymous mutations are generally selectively neutral or nearly so, in the case of a haploid organism such as a virus, we expect

$$\pi_S = 2N_e v \qquad \textbf{[equation 2.2]}$$

When we compare two populations of the same virus, we expect that v will probably be the same in the two populations. Therefore, comparing π_S in the two populations will provide an estimate of their relative effective population sizes.

In addition, the comparison of π_N and π_S provides information regarding the action of natural selection on the population of sequences under study. In most coding regions, π_S substantially exceeds π_N. This pattern occurs because most nonsynonymous mutations are deleterious and are therefore reduced in frequency or eliminated by purifying selection, whereas synonymous mutations are much more likely to be neutral or nearly neutral (Hughes 1999). The relative values of π_N and π_S are thus indicative of the strength and effectiveness of purifying selection. The strength of purifying selection reflects the functional importance of the protein or protein region being studied. In general, relative to π_S, we expect π_N to be lower in protein regions highly important to viral fitness than in protein regions that are less important to viral fitness.

When we have reason to suspect that positive Darwinian selection is acting to favor amino acid changes within a certain protein region, we may predict a reversal of the usual pattern, with π_N greater than π_S. An example of such a region would be a CD8+ TL epitope (Hughes et al. 2012); that is, a region of a viral protein that is recognized by a host class I major histocompatibility complex glycoprotein and presented to CD8+ T-lymphocytes ("cytotoxic T-cells"). In such a case, biological knowledge suggests a reason to expect repeated amino acid-altering changes in a region: namely, the evasion of the host immune system by the pathogen.

When there is no *a priori* reason to expect positive selection on some particular region of a viral protein, it may be useful to compute π_N and π_S in a sliding window along the gene. In the analysis of viruses infecting vertebrates, we frequently use a sliding window of 9 codons, because most CD8+ TL epitopes are nonamers (Evans et al. 1999; Hughes et al. 2001). Note that it is best to compute π_N and π_S separately, rather than to

compute the ratio π_N/π_S as is sometimes done. Ratios have undesirable statistical properties, and are therefore best avoided. For example, in the case of closely related sequences and a short sliding window length, π_S may often be zero in a given window, in which case the ratio π_N/π_S will be undefined. Additionally, examining the ratio π_N/π_S alone provides no information as to why that ratio is high in a given gene region. For example, the ratio π_N/π_S may be high in a certain region merely because π_S is unusually low, while π_N is not unusually high. In the latter case, high π_N/π_S would not be suggestive of positive selection but merely of some constraint on π_S, such as a low mutation rate or some constraint on synonymous substitution such as purifying selection on codon usage. Moreover, in the case of viruses, the existence of overlapping reading frames often provides constraints on synonymous substitutions because substitutions that are synonymous in one reading frame may be nonsynonymous in another (Hughes et al. 2001; Hughes and Hughes 2005).

2.3 Next-Generation Sequencing (NGS) Data

Nei and Kumar (2000) note that nucleotide diversity is equivalent to "heterozygosity at the nucleotide level." This relationship indicates that the estimation of nucleotide diversity across a genomic region does not require the availability of sequences (haplotypes) spanning the entire region, but rather only the frequency of different allelic variants at polymorphic sites. Thus, we can estimate nucleotide diversity from NGS data without reconstructing haplotypes, because NGS data provide information on the frequency of variants.

In order to estimate nucleotide diversity, we need to estimate the proportion of pairwise differences at each polymorphic site. Let m_i designate the coverage provided at the *i*th site, *i.e.*, the number of reads providing a base call for that site. The counts for the four bases at the *i*th site are designated, respectively, A_i, C_i, G_i, and T_i; thus $m_i = A_i + C_i + G_i + T_i$. The proportion of pairwise nucleotide differences at the *i*th site (D_i) is given by:

$$D_i = \frac{(A_i \times C_i)+(A_i \times G_i)+(A_i \times T_i)+(C_i \times G_i)+(C_i \times T_i)+(G_i \times T_i)}{(m_i{}^2 - m_i)/2} \quad \textbf{[equation 2.3]}$$

In within-host virus population data, the majority of SNPs are biallelic. In that case, only one of the six summed terms in the numerator of equation 2.3 will be non-zero. More complicated situations arise when multiple SNPs occur at the same site, or when analyses are based on entire codons. In the former case, there will be a maximum of 6 possible non-zero products. In the latter case, equation 2.3 must be expanded to compare all 64 possible codons, which constrains the number of non-zero terms in the numerator by an upper bound of ${}_{64}C_2 = 2{,}016$ pairs of codons.

In order to estimate nucleotide diversity in non-protein-coding regions (or without regard to coding differences in coding regions), for a sequence of L nucleotides and n polymorphic sites:

$$\pi = \sum_{i=1}^{n} \frac{D_i}{L} \quad \textbf{[equation 2.4]}$$

The same approach can be easily extended to estimate π_N and π_S in coding sequences. D_i is estimated separately for nonsynonymous and synonymous sites using equation 2.3,

while L represents the number of nonsynonymous or synonymous sites comprising the length of the sequence. This calculation obviously requires knowledge of a SNP's codon context. To compute synonymous D_i at a site that is less than fourfold degenerate, only the nucleotide pairs that are interchangeable without altering the amino acid are used in the numerator of equation 2.3. For example, consider the codon AAA, which encodes the amino acid Lys. If we are interested in determining π_N and π_S at this codon, we first note that one single-nucleotide variant at its third site is synonymous (AAG, also encoding Lys), while two single-nucleotide variants here are nonsynonymous (AAC and AAT, both encoding Asn). When estimating synonymous D_i at the third position of the codon, only the products which represent no amino acid change are used in the numerator of equation 2.3. Thus, in this case of AAA, the products used are A_i * G_i and C_i * T_i, representing the synonymous codon pairs AA**A**(Lys)/AA**G**(Lys) and AA**C**(Asn)/AA**T**(Asn), respectively. Conversely, to compute nonsynonymous D_i, only the other nucleotide pairs which *do* represent an amino acid change are included in the numerator of equation 2.3. In the case of AAA, the products used are A_i X C_i, A_i X T_i, C_i X G_i, and G_i X T_i, representing the nonsynonymous codon pairs AA**A**(Lys)/AA**C**(Asn), AA**A**(Lys)/AA**T**(Asn), AA**C**(Asn)/AA**G**(Lys), and AA**G**(Lys)/AA**T**(Asn), respectively. Thus, for the third position of the AAA codon, nonsynonymous D_i may be computed:

$$D_i = \frac{(A_i \times C_i) + (A_i \times T_i) + (C_i \times G_i) + (G_i \times T_i)}{(m_i^2 - m_i)/2} \quad \textbf{[equation 2.5]}$$

Similarly, synonymous D_i may be computed:

$$D_i = \frac{(A_i \times G_i) + (C_i \times T_i)}{(m_i{}^2 - m_i)/2} \quad \textbf{[equation 2.6]}$$

In the more complicated case of whole-codon comparisons, SNPs at multiple sites in the same codon are often present. The frequency of each possible codon in the population may be estimated using coverage information provided by NGS. All possible pairwise comparisons between codons (up to 2,016) are considered, contributing to π_N and π_S following the methods of Nei and Gojobori (1986).

The method we describe allows π_N and π_S to be calculated for pooled haploid NGS data. To automate this method, we have developed a software platform called SNPGenie (pronounced "snip genie"), which accepts SNP reports generated by separate SNP calling bioinformatics software (see Section 2.6). This approach differs from others, which estimate related population genetic parameters from aligned reads (*e.g.*, PoPoolation; Kofler et al. 2011; Raineri et al. 2012). The SNPGenie approach is flexible in that it can be easily modified to incorporate SNP reports generated using whatever is the preferred method for calling SNPs in pools. Thus our method takes advantage of the SNP calling software and settings that are most appropriate for the desired application. By separating the bioinformatics involved in SNP calling and evolutionary inference, our method allows more flexibility and ease in characterizing nucleotide diversity than has previously been possible. Additionally, unlike its predecessors, SNPGenie calculates: (1) d_N and d_S versus a reference sequence, characterizing divergence from an ancestral sequence; and (2) gene diversity at polymorphic sites, characterizing the magnitude and nature of synonymous, nonsynonymous, and ambiguous polymorphism. Finally, at a practical level, our method allows different quality measures (*e.g.*, filtering SNPs below a

minimum variant count) to be implemented without repeating the computationally intense process of SNP calling.

2.4 Example: Within-Host Diversity of SHFV

As an example of these methods, we present data on two new Arteriviruses isolated from natural populations of red colobus monkeys (*Procolobus rufomitratus tephrosceles*) from Uganda (Bailey et al. 2014). RNA was isolated from the blood plasma of wild-caught animals, and deep sequencing was performed on an Illumina MiSeq machine (Bailey et al. 2014). Many of the monkeys were infected by two distinct simian hemorrhagic fever viruses (SHFVs), designated SHFV-krc1 and SHFV-krc2. For 20 monkeys infected by both viruses, we estimated π_N and π_S for all codons with non-overlapping reading frames separately for the two viral genomes (Figure 2.1). Nucleotide diversity was consistently higher in the SHFV-krc1 virus than the SHFV-krc2 virus. Mean π_S for SHFV-krc1 was 0.0159 (± 0.00778 S.E.M.), which was significantly greater than that for SHFV-krc2 (0.00932 ± 0.00555; two-tailed $P = 0.00353$; paired T-test; Figure 2.1B). Mean π_N for SHFV-krc1 (0.00197 ± 0.000726) was also greater than that for SHFV-krc2 (0.00168 ± 0.000946), but this difference was not significant (two-tailed $P = 0.271$; paired T-test; Figure 2.1A). The hypothesis that purifying selection has acted to eliminate and/or to reduce the frequency of deleterious nonsynonymous mutations in these viruses was supported by the significantly lower mean π_N than mean π_S in each virus (two-tailed $P < 0.001$ in each case; paired T-tests).

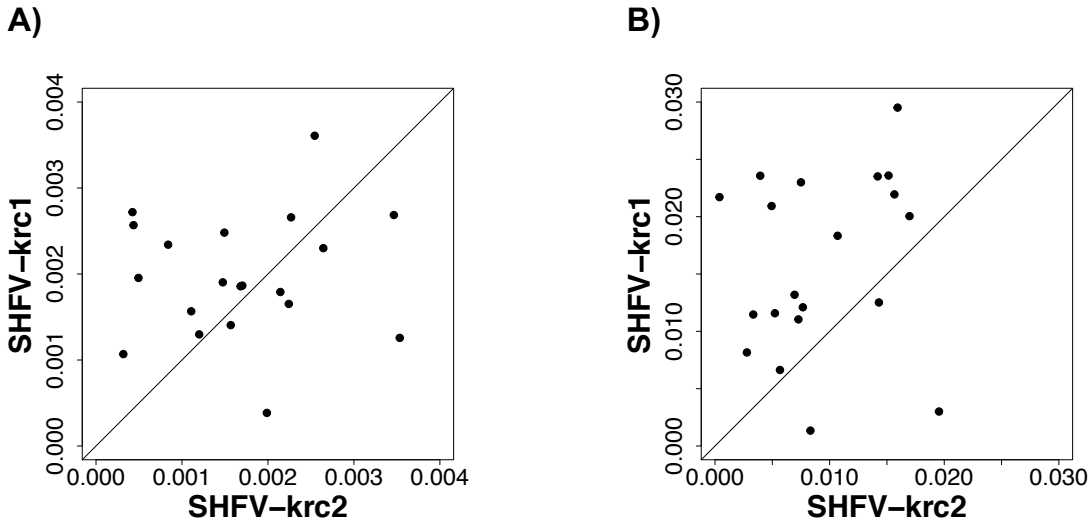


Figure 2.1. Nucleotide diversity in SHFV-krc1 vs. that in SHFV-krc2 from the same host red colobus monkey for all codons not overlapping multiple reading frames. Figures show (A) π_N and (B) π_S. In each case the line is a 45° line.

Population genetics theory predicts that neutral nucleotide diversity (reflected largely by π_S) is a function of both effective population size and mutation rate per generation (Nei 1987). SHFV-krc1 showed significantly greater viremia (blood concentration of virus) estimates than SHFV-krc2, suggesting the possibility that the within-host effective population sizes of SHFV-krc1 tend to be greater than those of SHFV-krc2 (Bailey et al. 2014), which would explain their difference in π_S (Figure 2.1B). On the other hand, it is also possible that the mutation rate per generation is higher in SHFV-krc1 than in SHFV-krc2. Preliminary analyses comparing of both viruses from one monkey sampled at two time points 2.5 years apart indeed suggested a higher mutation rate per unit time in SHFV-krc1 than in SHFV-krc2 (Bailey et al. 2014). However, the two viruses might still have identical mutation rates per generation if SHFV-krc1 has more generations per unit time. Resolving the relative contributions of

within-host effective population size, mutation rate, and generation time to the observed difference in nucleotide diversity in these two viruses will require further study (see Chapter 4).

2.5 Discussion

The population biology of viruses can be studied at two distinct levels: within hosts and across hosts. The study of within-host population biology of RNA viruses has been difficult until recently because it was necessary to infer features of a potentially very large and diverse viral population from only a small number of sequences. The availability of NGS methods that provide a much deeper picture of within-host viral diversity has a potential to change this situation dramatically. Using the methods described above we are able to obtain much more accurate estimates of nonsynonymous and synonymous nucleotide diversity than were previously possible, thereby providing insight into viral effective population sizes and the role of natural selection.

An aspect of the fundamental biology of viruses into which these methods may provide important insights revolves around the so-called "quasispecies theory," which models evolution in the case of infinite population sizes and high mutation rates (Eigen and Schuster 1977; Domingo 1992; Eigen 1996; Moya et al. 2000; Domingo 2002; Holmes and Moya 2002; Wilke 2005; Vignuzzi et al. 2006; Lauring and Andino 2010). Although there has been a tendency in the literature to treat quasispecies theory and population genetics as two competing paradigms, Holmes and Moya (2002) argue that the two might best be regarded as "two research traditions" each with its own "theoretical tools to explain population dynamics." Moreover, there are numerous overlaps between

quasispecies theory and traditional population genetics. Indeed, as Wilke (2005) has shown, the quasispecies model is mathematically equivalent to the mutation-selection balance model of classical population genetics.

Rather than contrasting quasispecies theory and population genetics as a whole, it might be more accurate to highlight differences between quasispecies theory and certain predictions of the neutral theory of molecular evolution (Kimura 1983). The original quasispecies models assumed infinite population sizes, as do the deterministic models of classical population genetics, although this obviously unrealistic assumption has been relaxed by some researchers working within the quasispecies tradition (*e.g.*, Park et al. 2010). On the other hand, the neutral theory emphasizes the importance of finite population size and genetic drift in the evolutionary process. As a consequence of genetic drift, populations are seen as inherently unstable and unpredictable in their genetic composition. By contrast, quasispecies theory tends to minimize the role of genetic drift and to predict the evolution of an equilibrium characterized by the dominance of a "cloud" of mutationally closely related genomes collectively known as a "quasispecies."

Empirical data that have been interpreted as providing support for quasispecies theory are often ambiguous and readily subject to alternative interpretations consistent with the neutral theory. For example, in experiments with laboratory-passaged strains of vesicular stomatosis virus (VSV), a strain with a high replication rate (and thus presumed high fitness) was outcompeted by a complex viral population assumed to represent a quasispecies (de la Torre and Holland 1990). However, this same result might be predicted under the neutral theory on the principle that, when the effective population size is low, natural selection is inefficient and even high-fitness genotypes may not

increase in frequency but rather may be subject to genetic drift (Kimura 1983). Since these virus populations were passaged (equivalent to "bottlenecking" in population genetic terms), they would be expected to have low effective population sizes (Hughes 2009).

Similarly, Lauring and Andino (2010) cite evidence that variants of dengue virus having a stop codon in one protein are maintained at high frequency in populations (Aaskov et al. 2006) as supporting a quasispecies model. But Aaskov et al. (2006) suggest other possible explanations for this observation that do not involve quasispecies. Since viruses in which certain proteins are defective can still be spread by "parasitizing" proteins from other viruses which co-infect the same host (Aaskov et al. 2006), selection against viruses with the stop codon may be relatively weak. Small effective population size, as a result of bottlenecks in transmission, may account for the failure of selection to remove such a mildly deleterious variant.

NGS methods can contribute to an increased understanding of within-host viral evolution, and thus to a resolution of some of the controversies raised by quasispecies theory. We will briefly discuss three types of relevant evidence to which NGS data and the aforementioned methods of analysis can contribute.

2.5.1 Nonsynonymous and Synonymous Polymorphism

Jenkins et al. (2001) have argued that a pattern whereby π_S exceeds π_N in VSV is evidence against the quasispecies theory because it implies that numerous synonymous mutations are neutral or nearly so, whereas the accumulation of neutral polymorphism is not predicted by the quasispecies model. However, the sequences which Jenkins et al.

(2001) analyzed were sampled from numerous different hosts; thus, because they did not represent within-host populations, the relevance of these data to the quasispecies model of within-host virus evolution might be questioned. Sanger sequencing of within-host populations of viruses has shown a pattern whereby π_S substantially exceeds π_N in a variety of viruses (Hughes, Piontkivska, et al. 2005; Callendret et al. 2011; Li et al. 2011). Similar patterns have been seen in studies using NGS data (Hughes et al. 2012; Lauck et al. 2012; Wilker et al. 2013; Bailey et al. 2014). Further studies using NGS methods will make it possible to estimate the relative magnitude of nonsynonymous and synonymous polymorphism for within-host virus populations, and thus to assess the role of neutral mutations and genetic drift in within-host viral evolution.

2.5.2 Increase in Polymorphism Over Time

The neutral theory predicts that most polymorphism in natural populations is selectively neutral or nearly so. Thus, in the absence of perturbing factors such as radical changes in the selective regime or population bottlenecks, neutral polymorphism will accumulate over time as a consequence of mutation. The quasispecies theory, by contrast, predicts that an equilibrium state will develop after which polymorphism will not increase. So far relatively few studies have examined within-host viral polymorphism at several time points over the course of infection; however, several studies using Sanger sequencing (Callendret et al. 2011; Li et al. 2011) have provided evidence that polymorphism—particularly synonymous polymorphism—increases over time, as predicted by the neutral theory. Particularly interesting were data showing a steady increase over time of within-host viral π_S in human patients, raging from 2 to 38 years post-infection with hepatitis C

virus (Li et al. 2011). It is important to test for the generality of this pattern across different RNA virus species. Because NGS methods provide the potential for examining genome-wide viral polymorphism at different time points over the course of infection, these methods seem particularly well designed for addressing this question.

2.5.3 The Impact of Effective Population Size

According to the neutral theory, the extent of sequence polymorphism maintained in a population should be correlated with its effective population size, while quasispecies theory argues that within-host populations of RNA viruses are so large that effective population size can be ignored. Results such as those of Bailey et al. (2014) support the neutral theory since they suggest a correlation between nucleotide diversity and viral load (viremia), which may reflect viral population size. The correlation between virus nucleotide diversity and viral load requires further testing in a variety of viruses.

In addition to the potential utility of NGS analyses in addressing theoretical debates regarding quasispecies theory, the approaches described here are useful in studying a number of other questions regarding within-host virus evolution. They can provide evidence regarding positive selection favoring new viral mutants, including those that confer escape from host immune recognition mechanisms (Hughes et al. 2012); those that confer resistance to anti-viral drugs; and those that are favored because they better adapt the virus to a new host species (Wilker et al. 2013).

2.6 Note on Software

In order to perform the analyses herein, we developed and implemented a nascent software platform called SNPGenie (Wilker et al. 2013; Bailey et al. 2014) for analyzing nonsynonymous and synonymous polymorphism in these pooled NGS samples. SNPGenie makes several advances over previous approaches (Kofler, Orozco-terWengel, et al. 2011; Raineri et al. 2012), and is described in Chapter 3.

2.7 Acknowledgments and Funding

This research was supported by United State National Institutes of Health (NIH) grant AI077376 to David H. O'Connor and A.L.H.; by NIH grant AI096882 to Jonathan Honegger, Christopher Walker, and A.L.H.; and by United States National Science Foundation Graduate Research Fellowship DGE-0929297 and a University of South Carolina Presidential Fellowship to C.W.N. No conflicts of interest declared.

CHAPTER 3

SNPGENIE: ESTIMATING EVOLUTIONARY PARAMETERS TO DETECT NATURAL SELECTION USING POOLED NEXT-GENERATION SEQUENCING DATA[2]

[2] Nelson CW, Moncla LH, Hughes AL. 2015. SNPGenie: estimating evolutionary parameters to detect natural selection using pooled next-generation sequencing data. *Bioinformatics* 31(22):3709-11; https://github.com/hugheslab/snpgenie. Reprinted here in modified form with permission of publisher; see Appendix A.

3.1 Introduction

Next-generation sequencing (NGS) technologies allow the rapid sequencing of pooled DNA samples containing genetic material from multiple individuals. The resultant single-nucleotide polymorphism (SNP) data may be used to reliably estimate population genetic parameters with more accuracy and less expense than the separate sequencing of multiple individuals (Futschik and Schlötterer 2010; Lynch et al. 2014), especially when samples are large and coverage is high. Unfortunately, high coverage data also suffer from a substantial false-positive error rate. SNP calling techniques can address this issue, but the only software currently available for evolutionary analysis of pooled NGS data, PoPoolation (Kofler, Orozco-terWengel, et al. 2011), is inextricable from a problematic SNP caller that has an extremely high false-positive rate (Raineri et al. 2012). Further, PoPoolation relies on large pileup files and problematic simplifications, including use of the reference sequence alone to determine the number of nonsynonymous and synonymous sites. Ideally, software for evolutionary analyses of these data would allow users to first call SNPs using any preferred method, and then process the results using standard methods for determining the numbers of nonsynonymous and synonymous differences and sites.

We have developed SNPGenie to meet this need, available at https://github.com/hugheslab/snpgenie. Using SNP calling results, SNPGenie estimates: (i) nucleotide diversity (π), and its nonsynonymous and synonymous partitions (π_N and π_S, respectively) for coding regions; (ii) mean nonsynonymous and synonymous divergence (d_N and d_S, respectively) from a reference sequence; (iii) gene diversity (H; Nei 1987); (iv) site type classification (nonsynonymous, synonymous or ambiguous) for

polymorphic coding loci (Knapp and Hughes 2012); and (v) the constraint imposed by overlapping open reading frames. These parameters do not depend on linkage (see Chapter 2), circumventing a major limitation of pooled data for other applications (Cutler and Jensen 2010). Indefinitely large genomes with multi-nucleotide variants may be analyzed at speeds exceeding those of PoPoolation's default settings. The results allow users to test evolutionary hypotheses on the roles of negative (purifying) selection, positive (Darwinian) selection, and random genetic drift in the sampled population. In general, $\pi_N = \pi_S$ indicates neutrality, $\pi_N < \pi_S$ indicates purifying selection, and $\pi_N > \pi_S$ may indicate positive selection favoring multiple amino acid changes (Hughes 1999). Comparing *H* at distinct polymorphic site categories may also address these hypotheses (although an important limitation is explored in Chapter 4) (Hughes et al. 2003). Parameter estimates are available at the nucleotide, codon, sliding window, whole gene and whole genome/population levels.

3.2 Methods

SNPGenie is a data processing Perl script, with no additional dependencies. The program accepts a reference sequence(s) (FASTA), a Gene Transfer Format (GTF) file with CDS annotations, and an arbitrary number of SNP reports, currently including Geneious (Variations/SNPs Annotations Table), CLC Genomics Workbench (Annotated Variant File), and VCF (variant call format) formats. The download package includes the Perl code, a detailed manual (README), various example files, and scripts to aid in data preparation.

Nucleotide diversity (π) is the mean number of pairwise differences per site in a population of sequences. SNPGenie estimates this for all sites, and then separately for nonsynonymous and synonymous coding sites (π_N and π_S, respectively), using a new method (Chapter 2) based on that of Nei and Gojobori (1986). Differences are calculated using all comparisons within every polymorphic codon and including all mutational pathways. To calculate the numbers of nonsynonymous and synonymous sites, SNPGenie weights by the sample allele frequencies. This becomes especially important when populations diverge from the reference sequence(s). Gene diversity is calculated as $H = 1 - \sum x_i^2$, where x_i is the population frequency of nucleotide i. Polymorphic coding sites are classified following the methods of Knapp and Hughes (2012), with gene diversities given for each category.

SNPGenie (version 1.2) and PoPoolation (version 1.2.2) were used to analyze pooled H5N1 data from ferret #3501-DPI5, obtained from Wilker et al. (2013) (Jorge Dinis, personal correspondence). SNPGenie used SNP calling results from Geneious Version 5.6.3, while PoPoolation necessarily performed its own SNP calling. For SNPGenie, all default values were used. For PoPoolation, the `Syn-nonsyn-sliding.pl` script was used with default settings, except `max-coverage` = 100000, `dissable-corrections` = on, `min-count` = 1, `window-size` = 3, and `step-size` = 3 (single codon analysis). Statistical analyses were performed using RStudio version 0.98.1049.

Table 3.1. Mean nonsynonymous (N) and synonymous (S) differences and sites in hemagglutinin (HA) and neuraminidase (NA) genes of an H5N1 influenza population, estimated by PoPoolation and SNPGenie.

Gene	Parameter	R^2	PoPoolation	SNPGenie	*P*
HA	N diffs	0.991	0.0039 ± 0.0015	0.0033 ± 0.0016	0.001
	S diffs	0.995	0.0011 ± 0.0006	0.0008 ± 0.0006	0.002
	N sites	0.830	2.3844 ± 0.0144	2.3483 ± 0.0144	< 0.001
	S sites	0.830	0.6156 ± 0.0144	0.6517 ± 0.0144	< 0.001
NA	N diffs	0.437	0.0015 ± 0.0001	0.0089 ± 0.0007	< 0.001
	S diffs	0.231	0.0007 ± 0.0001	0.0029 ± 0.0002	< 0.001
	N sites	0.882	2.3667 ± 0.0159	2.3347 ± 0.0158	< 0.001
	S sites	0.884	0.6333 ± 0.0159	0.6647 ± 0.0158	< 0.001

Values shown are means ± standard errors. *P*-values refer to a paired T-test comparing PoPoolation and SNPGenie, with the codon as the unit. For all R^2, $P < 0.001$ (F test).

3.3 Results

To validate SNPGenie's execution of the Nei-Gojobori (1986) method, we constructed sequences with all 61 non-STOP codons and known numbers of differences. MEGA Version 6 (Tamura et al. 2013) was used to calculate π_N and π_S. SNP reports and GTF files were then constructed to reflect the known variant frequencies and reference sequence, and SNPGenie was used to estimate the same parameters. All results agreed to the last decimal.

Next, both SNPGenie and PoPoolation were used to analyze a pooled H5N1 sample. The nonsynonymous and synonymous mean numbers of pairwise differences per site and numbers of sites (the numerator and denominator of π_N and π_S) were then estimated for the hemagglutinin (HA) and neuraminidase (NA) genes.

When PoPoolation estimates were regressed on those from SNPGenie, all R^2 values were significant ($P < 0.001$; F-test), but smaller for differences in NA. PoPoolation overestimated differences for HA and underestimated them for NA, while overestimating the number of nonsynonymous sites (Table 3.1). π was significantly lower in HA ($P < 0.01$ for π_N; $P < 0.001$ for π_S; two-sample T-tests), consistent with previous evidence for a population bottleneck upon viral transmission that is driven by selection for specific HA residues (Wilker et al. 2013). Because PoPoolation overestimated differences in HA, this suggests that its false discovery rate may be exacerbated in low-diversity (*e.g.*, bottlenecked) contexts.

Most differences between SNPGenie and PoPoolation can be attributed to: (i) differences in SNP calling; (ii) PoPoolation's treatment of STOP codon variants as nonsynonymous; and (iii) SNPGenie's use of allele frequency data in determining the number of sites, contrasted to PoPoolation's use of the reference sequence alone. PoPoolation also reports $\pi_S = 0$ for codons with no synonymous sites, where π_S should be undefined. This could highly inflate the π_N/π_S ratio, overestimating the prevalence of positive natural selection. If the false positive calls are random, ~75% will be nonsynonymous (Graur and Li, 2000), exacerbating this problem.

Planned future improvements in SNPGenie include additional SNP report formats (*e.g.*, VCF) and weighted mutational pathways.

3.4 Software Operation and Use

A brief description of SNPGenie use follows. Its basic applications are all accomplished using the main script, `snpgenie-1.2.2.pl`, in a directory containing the necessary input files. Other accessory scripts currently available are described in Section 3.4.5.

3.4.1 SNPGenie Input

SNPGenie version 1.2 is a command-line interface application written in Perl, with no additional dependencies. As such, it is limited only by the memory and processing capabilities of the local hardware. As input, it accepts:

1. One or more reference sequence files in FASTA format (.fa/.fasta);
2. One file with CDS information in Gene Transfer Format (.gtf); and
3. One or more tab-delimited (.txt) SNP reports in CLC, Geneious, or VCF format.

For ease and simplicity, one need only run SNPGenie in a directory containing the necessary input files, and SNPGenie takes care of all processing (Section 3.4.2 describes options for more control). To do this, the user first downloads the `snpgenie-1.2.2.pl` script and places it in the system's `PATH`, or simply in the working directory. Next, the SNP report(s), FASTA(s) (.fa/.fasta), and GTF (.gtf) files are placed in the working directory. The command line prompt (or Terminal) is used to navigate to the directory containing these files using the `cd` command. Finally, SNPGenie is executed by typing the name of the script and pressing the `<RETURN>` (or `<ENTER>`) key. Further details on input are given below.

3.4.1.1 Reference Sequence

Only one reference sequence may be provided in a single FASTA (.fa/.fasta) file. Thus, all SNP coordinates in the SNP report(s) should have been called relative to the single reference sequence. This ONE-SEQUENCE MODE allows the maximum number of estimations to be performed, and is the only mode of SNPGenie that remains supported. (A MULTI-SEQUENCE MODE was available in past versions.) Because of this one-sequence stipulation, a script has been provided to split a multi-sequence FASTA file into its constituent sequences if need be; see Section 3.4.5.

3.4.1.2 Gene Transfer Format

The Gene Transfer Format (.gtf) file is tab (`\t`)-delimited, and must include non-redundant records for all CDS elements (*i.e.*, open reading frames, or ORFs) present in the SNP report(s). Note that SNPGenie expects every coding element to be labeled as type "CDS", and for its product name to follow a "gene_id" tag. In the case of CLC and Geneious SNP reports, this name must match that present in the SNP report. If a single coding element has multiple segments (*e.g.*, exons) with different coordinates, the user simply enters one line for each segment, using the same product name. (Although SNPGenie could only handle 2 segments per ORF in the past, there is now no limit.) Finally, for cases with reverse '-' strand features, SNPGenie must be run twice, once for each strand, with that strand's own set of input files (*i.e.*, the '-' strand FASTA, GTF, and SNP report); see Section 3.4.1.4. The Brent Lab provides more information about GTF at http://mblab.wustl.edu/GTF22.html. A simple example follows:

```
reference.gbk   CLC CDS 5694    8369    .   +   0   gene_id "ORF1";
reference.gbk   CLC CDS 8203    8772    .   +   0   gene_id "ORF2";
reference.gbk   CLC CDS 1465    4485    .   +   0   gene_id "ORF3";
reference.gbk   CLC CDS 5621    5687    .   +   0   gene_id "ORF4";
reference.gbk   CLC CDS 7920    8167    .   +   0   gene_id "ORF4";
reference.gbk   CLC CDS 5395    5687    .   +   0   gene_id "ORF5";
reference.gbk   CLC CDS 7920    8016    .   +   0   gene_id "ORF5";
reference.gbk   CLC CDS 4439    5080    .   +   0   gene_id "ORF6";
reference.gbk   CLC CDS 5247    5549    .   +   0   gene_id "ORF7";
reference.gbk   CLC CDS 4911    5246    .   +   0   gene_id "ORF8";
```

3.4.1.3 SNP Reports

Each SNP report submitted to SNPGenie should contain variant calls for a single pooled-sequencing run (*i.e.*, a single population) in one of the following formats:

- CLC Genomics Workbench. At minimum, the CLC Genomics Workbench SNP report must include the following default column selections, with the unaltered CLC column headers:
    - **Reference Position**, which refers to the start site of the polymorphism within the reference FASTA sequence;
    - **Type**, which refers to the nature of the record, usually the type of polymorphism, *e.g.*, "SNV" for single-nucleotide variants;
    - **Reference**, the reference nucleotide(s) at that site(s);
    - **Allele**, the variant nucleotide(s) at that site(s);
    - **Count**, the number of reads containing the variant;
    - **Coverage**, the total number of sequencing reads at the site(s);
    - **Frequency**, the frequency of the variant as a percentage, *e.g.*, "14.6" for 14.60%; and
    - **Overlapping annotations**, containing the name of the protein product or open reading frame (ORF), *e.g.*, "CDS: ORF1".

In addition to the aforementioned columns, the SNP report should ideally be free of thousands separators (,) in the **Reference Position**, **Count**, and **Coverage** columns (default format). The **Frequency** must remain a percentage (default format). Finally, the user should verify that the reading frame in the CLC output is correct. SNPGenie will produce various errors to indicate when these conditions are not met, *e.g.*, by checking that all products begin with START and end with STOP codons, and checking for premature stop codons. Relevant information will be printed to the SNPGenie LOG file.

- Geneious. At minimum, the Geneious SNP report must include the following default column selections, with the unaltered Geneious column headers:
    - **Minimum** and **Maximum**, which refer to the start and end sites of the polymorphism within the reference FASTA sequence, and will hold the same value for SNP records;
    - **CDS Position**, with the coordinate of the site relative to the start cite of the relevant CDS annotation;
    - **Type**, which refers to the nature of the record entry, *e.g.*, "Polymorphism";
    - **Polymorphism Type**, which gives the type of polymorphism;
    - **product**, containing the name of the protein product or open reading frame, *e.g.*, ORF1;

- **Change**, which contains the reference and variant nucleotides, *e.g.*, "A -> G", and are always populated for SNP records;
- **Coverage**, containing the number of sequencing reads that include the site; and
- **Variant Frequency**, which contains the frequency of the nucleotide variant as a percentage, *e.g.*, 14.60%.

As with CLC, the Geneious SNP report should ideally be free of extraneous characters such as thousands separators (,), but SNPGenie will do its best to adapt if they are present. Again, the **Variant Frequency** must remain a percentage (default format); again, the user should verify that the reading frame in the Geneious output is correct. SNPGenie will produce various errors to indicate when these conditions are not met, *e.g.*, by checking that all products begin with START and end with STOP codons, and checking for premature stop codons. Relevant information will be printed to the SNPGenie LOG file.

- Variant Call Format (VCF). At minimum, the VCF SNP report must include (and at present does so by definition) the following columns, with the unaltered VCF column headers:
  - **CHROM**, the name of the reference genome;
  - **POS**, which refers to the start site of the polymorphism within the reference FASTA sequence;
  - **REF**, the reference nucleotide(s) at that site(s);
  - **ALT**, the variant nucleotide(s) at that site(s);

- **QUAL**, the Phred quality score for the variant;
- **FILTER**, the filter status, based on such metrics as minimum frequencies and minimum quality scores;
- **INFO**, additional necessary information, including entries for:
    - If a pooled VCF (*i.e.*, the SNPs are called from a pooled sequencing sample):
        - **DP4**, containing the number of reference and variant reads on the forward and reverse strands (*e.g.*, "DP4=11,9,219,38")
    - If a summary VCF (*i.e.*, the SNPs from multiple individual sequencing samples are being summarized):
        - **NS**, the number of samples (*i.e.*, individual sequencing experiments) being summarized; and
        - **AF**, the allele frequency(-ies) for the variant alleles in the same order as listed in the ALT column (*e.g.*, "NS=30" and "AF=0.200")
- **FORMAT** and **SAMPLE** as an alternative to **INFO** for the pooled VCF approach (*i.e.*, the SNPs are called from a pooled sequencing sample), with data entries for:
    - **AD**, the allele depth for the reference, followed by that for the variant allele(s) in the same order as listed in the **ALT** column (*e.g.*, "AD" in the **FORMAT** column and "75,77" in the **SAMPLE** column); and **DP**, the coverage or total read depth (*e.g.*,

"DP" in the **FORMAT** column and "152" in the **SAMPLE** column)

As usual, the user must make sure to maintain the VCF file's features, such as TAB-delimited columns. Unlike some other formats, the allele frequency in VCF is a decimal.

3.4.1.4 A Note on Reverse Complement ('-' Strand) Records

Many large genomes have coding products on both strands. In this case, SNPGenie must be run twice: once for the '+' strand, and once for the '-' strand. This requires FASTA, GTF, and SNP report input for the '-' strand. The script `snpgenie-vcf2revcom.pl`, described in Section 3.4.5, automatically creates these files for the user, using the original data. Note that, regardless of the original SNP report format, the reverse complement SNP report is in a CLC-like format that SNPGenie will recognize. For both runs, the GTF should include all products for both strands, with the products on the strand being analyzed classified as '+' and having coordinates defined with reference to the beginning of that FASTA sequence. Also note that a GTF file containing only '-' strand records will not run; SNPGenie does calculations only for the products on the current '+' strand, using the '-' strand products only to determine the presence of overlapping reading frames.

3.4.2 Options

In case the user wishes to alter the way SNPGenie works, the following options (implemented using Perl's `Getopt::Long` module) may be used:

- `--minfreq`: optional floating point parameter specifying the minimum allele (SNP) frequency to include. Entered as a proportion/decimal (*e.g.*, 0.01), not as a percentage (*e.g.*, not 1.0%). Default: 0.
- `--snpreport`: optional string parameter specifying the (one) SNP report to analyze. Default: auto-detect `.txt` and `.csv` file(s).
- `--fastafile`: optional string parameter specifying the (one) reference sequence. Default: auto-detect `.fa` and/or `.fasta` file(s).
- `--gtffile`: optional string parameter specifying the one file with CDS annotations. Default: auto-detect the `.gtf` file.
- `--sepfiles`: optional Boolean (flag) parameter specifying whether to produce separate results (codon) files for each SNP report (all results already included together in the `codon_results.txt` file). Simply include in the command line to activate. Default: not included.
- `--slidingwindow`: optional integer parameter specifying the length of the sliding (codon) window used in the analysis. Default: 9 codons.
- `--ratiomode`: optional Boolean (flag) parameter specifying whether to include π values for each codon in the `codon_results.txt` file(s). This is usually inadvisable, as π values (especially π_S) are subject to great stochastic error. Simply include in the command line to activate. Default: not included.
- `--sitebasedmode`: optional Boolean (flag) parameter specifying whether to include π values derived using a site-based (reference codon context only) approach in the `codon_results.txt` file(s). This is usually inadvisable, as π

values will not reflect the true population pairwise comparisons. Simply include in the command line to activate. Default: not included.

For example, if the user wishes to activate the `sepfiles` option, specify a minimum allele frequency of 1%, and specify input files, they might enter the command:

```
snpgenie-1.2.2.pl –sepfiles --minfreq=0.01 --snpreport=mySNPreport.txt
      --fastafile=myFASTA.fa --gtffile=myGTF.gtf
```

3.4.3 How SNPGenie Works

Given the appropriate files, SNPGenie calculates gene and nucleotide diversities for different types of sites in a protein-coding sequence. Nucleotide diversity may be defined as the average number of nucleotide variants per nucleotide site for all pairwise comparisons. To distinguish between nonsynonymous and synonymous differences and sites, it is necessary to consider the codon context of each nucleotide in a sequence. This is why the user must submit the starting and ending sites of the coding regions in the `.gtf` file, along with the reference FASTA sequence file, so that the numbers of nonsynonymous and synonymous sites for each codon may be accurately estimated by the Nei-Gojobori (1986) method. SNPGenie first splits the coding sequence into codons, each of which contains 3 nucleotide sites. The software then determines the numbers of these sites which are nonsynonymous and synonymous by testing all polymorphisms present at each site of every codon in the sequence. Because different nucleotide variants at the same site may lead to both nonsynonymous and synonymous polymorphisms, fractional sites occur frequently (*e.g.*, only 2 of 3 possible nucleotide substitutions at the

third position of AGA cause an amino acid change; thus, that site is considered 2/3 nonsynonymous and 1/3 synonymous). Next, the SNP report is consulted for the presence of variants to produce a revised estimate. Variants are incorporated through averaging weighted by their frequency. Although it is relatively rare, high levels of sequence variation may alter the number of nonsynonymous and synonymous sites in a particular codon, contributing to an altered picture of natural selection.

Next, SNPGenie calculates the number of nucleotide differences for each codon in each ORF specified in the `.gtf` file. Calculating nucleotide diversity codon-by-codon enables sliding window analyses that may help to pinpoint important nucleotide regions subject to varying forms of natural selection. SNPGenie determines the average number of pairwise differences as follows: for every variant in the SNP Report, the number of variants is calculated as the product of the variant's relative frequency and the coverage at that site. For each variant nucleotide (up to 3 non-reference nucleotides), the number of variants is stored, and their sum is subtracted from the coverage to yield the reference's absolute frequency. Next, for each pairwise nucleotide comparison at the site, it is determined whether the comparison represents a nonsynonymous or synonymous change. If the former, the product of their absolute frequencies contributes to the number of nonsynonymous pairwise differences; if the latter, it contributes to the number of synonymous pairwise differences. When comparing codons with more than one nucleotide difference, all possible mutational pathways are considered, per the method of Nei and Gojobori (1986). The sum of pairwise differences is divided by the total number of pairwise comparisons at the codon (${}_n\mathrm{C}_2$, where n is coverage) to yield the mean

number of differences per site of each type. This is calculated separately for nonsynonymous and synonymous comparisons. For further background, see Chapter 2.

3.4.4 Output

SNPGenie creates a new folder called `SNPGenie_Results` within the working directory. This contains the following TAB-delimited results files, for which detailed documentation can be found at https://github.com/hugheslab/snpgenie:

1. `SNPGenie_parameters.txt`, containing the input parameters and file names.
2. `SNPGenie_LOG.txt`, documenting any peculiarities or errors encountered. Warnings are also printed to the Terminal (shell) window.
3. `site_results.txt`, providing results for all polymorphic sites. Note that, if the population is genetically homogenous at a site, even if it differs from the reference or ancestral sequence, it will not be considered polymorphic. Also keep in mind that columns are sorted by product first, then site number, with noncoding sites at the end of the file.
4. `codon_results.txt`, providing results for all codons.
5. <SNP report name(s)>`_results.txt`, containing the information present in the `codon_results.txt` file, but subset by SNP report.
6. `product_results.txt`, providing summary results for all CDS elements present in the GTF file for the '+' strand.
7. `population_summary.txt`, providing summary results for each population's sample (SNP report) with respect to the '+' strand.

8. `sliding_window_length_results.txt`, containing codon-based results over a sliding window, with a default length of 9 codons.

3.4.5 Additional Scripts

Some additional scripts are included to automate some common tasks when preparing SNPGenie input. These currently are:

- `snpgenie-gbk2gtf.pl`. At the command line, this script is provided with one argument: a GenBank (.gbk) file. It will extract the coding element annotations to produce a Gene Transfer Format (.gtf) file ready for SNPGenie. Here's an example:

```
snpgenie-gbk2gtf.pl my_genbank_file.gbk
```

- `snpgenie-gff2gtf.pl`. At the command line, this script is provided with one argument: a General Feature Format (.gff) file. It will extract the coding element annotations to produce a Gene Transfer Format (.gtf) file ready for SNPGenie, with “gene_id” annotations identified using the GFF “ID” tag. Here's an example:

```
snpgenie-gbk2gtf.pl my_gff_file.gff
```

- `snpgenie-split_fasta.pl`. At the command line, this script is provided with one argument: a FASTA (.fa or .fasta) file containing multiple sequences. It

will create multiple files in the working directory, each containing one of the sequences. Here's an example:

```
snpgenie-split_fasta.pl my_multi_fasta_file.fasta
```

- `snpgenie-vcf2revcom.pl`. This script automates the creation of the reverse complement input files. At the command line, it is provided with three arguments, in the following order:
    i. A '+' strand FASTA (`.fa` or `.fasta`) file containing the reference sequence against which SNPs were called;
    ii. A '+' strand GTF file containing both '+' and '–' strand products from the '+' strand point of view; and
    iii. A '+' strand SNP report in VCF format.

This script will then create a '-' strand (reverse complement) version of each file in the working directory, with "`_revcom`" concatenated to the original file name. Here's an example:

```
snpgenie-vcf2revcom.pl my_snp_report.vcf
     my_reference_sequence.fasta my_cds_file.gtf
```

3.4.6 Studies Using SNPGenie

To date, SNPGenie has been used to study H5N1 (Wilker et al. 2013) and H1N1 (Moncla et al. 2016) influenza; simian hemorrhagic fever virus (Bailey et al. 2014; Nelson and

Hughes 2015; Chapters 2 and 4), simian immunodeficiency virus (SIV) (Gellerup et al. 2016); Arteriviruses, pegiviruses, and lentiviruses in African Green Monkeys (Bailey et al., in press at *Journal of Virology*). Its first application to a eukaryote, namely the Nod-Like Receptor resistance genes of the wild tomato *Solanum pennelli*, is in press at *Genome Biology and Evolution* (Stam et al.).

3.5 Conclusion

I have developed a software tool capable of performing population genetic analyses with numerous forms of pooled-sequencing (and other) NGS variant data. This affords an opportunity in Chapter 4 to return to the data of Bailey et al. (2014) to test numerous predictions about within-host viral evolution, including those presented in Chapter 2.

3.6 Acknowledgments and Funding

This research was supported by NIH grant Al077376 to David H. O'Connor (D.H.O.) and A.L.H.; by NIH grant Al096882 to Christopher Walker and A.L.H; by NIH R01 grant Al084787 to D.H.O. and Thomas C. Friedrich (L.H.M.); by UW-Madison Molecular Biosciences Training Grant T32 GM07215 to L.H.M.; and by an NSF GRF (DGE-0929297), a University of South Carolina (USC) Presidential Fellowship, and a USC Dept. Biol. Sci. Kathryn Hinnant-Johnson, M.D. Memorial Fellowship to C.W.N. No conflicts of interest declared.

CHAPTER 4

POOLED SEQUENCING VIRAL DATA ALLOW EPITOPE DISCOVERY, EVOLUTIONARY MODELING, AND MUTATION RATE ESTIMATION: PROSPECTS AND LIMITATIONS WITH RED COLOBUS ARTERIVIRUSES[3]

[3] Nelson CW, Bailey AL, Lauck M, Dinis JM, Sibley SD, Goldberg TL, O'Connor DH, Hughes AL. In preparation for *Molecular Biology and Evolution*.

4.1 Introduction

RNA viruses exhibit several properties that increase their likelihood of emergence via host-switching as compared to DNA viruses, most notably high replication and mutation rates (Holmes 2009; Chapter 1). As a result, they are not greatly limited by factors that may restrict the rate of adaptive evolution in other systems with low reproduction rates (Haldane 1957; Nelson 2015), making them a particularly tractable model for studying the role of selection in evolution. The *Arteriviridae* family (Arteriviruses) is a group of positive-sense, single-stranded RNA viruses that infect mammals, causing both persistent asymptomatic infections and acute disease, depending on the host (Snijder et al. 2013).

Lauck et al. (2011) recently discovered two closely related Arteriviruses infecting 30 red colobus (RC) monkeys (*Procolobus rufomitratus tephrosceles*) in Uganda's Kibale National Park: simian hemorrhagic fever viruses (SHFVs) krc1 and krc2. SHFV-krc1 and SHFV-krc2 (hereafter krc1 and krc2) each have genomes approximately 15,500 nt in length and share 14 open reading frames (ORFs). We adopt the nomenclature of Snijder et al. (2013), with ORFs ordered from 5' to 3' by start site as 1a, TF, 1b, 2a', 2b', 3', 4', 2a, 2b, 3, 4, 5a, 5, 6, and 7 (Figure 4.1). Ancient duplications have likely given rise to the ORF pairs 2a'/2a, 2b'/2b, 3'/3, and 4'/4 (Godeny et al. 1998). Among the Arteriviruses, ORFs 2a', 3', and 4' are unique to SHFVs (formerly ORFs 2a, 2b, and 3; Lauck et al. 2013). Note that Lauck et al. (2011) use a different naming system which lacks our ORFs TF, 2b', and 5a. Both viruses appear to be asymptomatic in RC hosts and, consistent with patterns expected for viruses likely to emerge by host-switching, they exhibit high prevalence, viremia, genetic diversity, and evolutionary substitution rates.

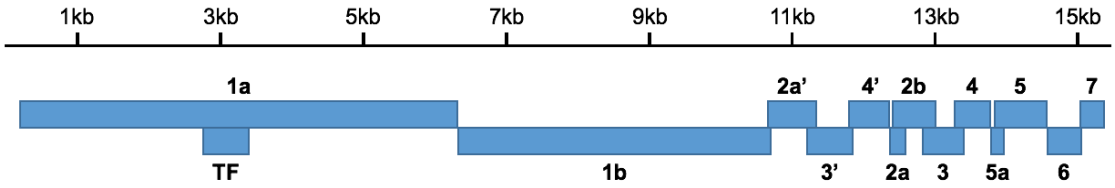


Figure 4.1. The SHFV genome. Both krc1 and krc2 are positive-sense single-stranded RNA viruses that share the same open reading frames (ORFs) and genome length (approximately 15,500 nt), with minor differences (Table 4.1). ORFs are offset to show their degree of overlap with one another, and we use the nomenclature of Snijder et al. (2013). ORF TF is contained entirely within 1a, and all other ORFs overlap at their termini. Prime (') symbols indicate an ancient duplication, such that ORF 2a' is thought to be a duplicate of 2a, and so on. The first three ORFs make up well over half the genome.

Population genetic parameters such as nucleotide diversity (π), equivalent to the average number of pairwise differences per site in a population (Li 1997), can shed light on the within- and between-host evolutionary dynamics of infectious agents. For example, π can be estimated separately for nonsynonymous (amino acid-altering) and synonymous (silent) nucleotide sites in coding regions, yielding π_N and π_S, respectively (Chapter 2). In the case of within-host viral populations, values of π are generally quite low (<<10%), allowing the use of simple estimation methods (Nei and Gojobori 1986; Nei and Kumar 2000). Because nonsynonymous mutations introduce amino acid changes, they are far more likely than synonymous changes to alter fitness by disrupting protein structure. As a result, purifying (negative) selection generally acts to decrease the frequencies of nonsynonymous mutations in the population. On the other hand, synonymous mutations have relatively "silent" effects, as evidenced by their preponderance in nearly all populations (Hughes 1999), including all viruses (Holmes 2009). As a result, most synonymous mutations accumulate freely, as if neutral or nearly neutral. Thus, in instances where purifying selection acts to eliminate nonsynonymous

mutations, we expect $\pi_N < \pi_S$. When purifying selection is relaxed, both types of sites evolve neutrally at the same rate, and we expect $\pi_N = \pi_S$. Finally, when positive (Darwinian) selection acts to promote repeated amino acid changes, as is the case with overdominant positive selection (heterozygote advantage), we expect $\pi_N > \pi_S$.

Instances of $\pi_N > \pi_S$ are not unambiguous evidence of overdominant selection, let alone other forms of positive selection. For example, $\pi_N > \pi_S$ can result from other situations, *e.g.*, an increase in the frequency of rare variant alleles during a population bottleneck. Nor is selection always expected to produce this signal, *e.g.*, the fixation of variant alleles, resulting in $\pi = 0$. However, in the case of host/pathogen coevolution, the biological context allows the *a priori* hypothesis that the host immune system will target the pathogen. This is expected to provide a selective pressure that promotes mutations in regions of its genome encoding antigenic peptides (epitopes) allowing immune escape.

Several studies have introduced statistical approaches for estimating population genetic parameters from newly available pooled next-generation sequencing (NGS) variant data, *i.e.*, the simultaneous sequencing of multiple individuals in a single sample (Kofler, Orozco-terWengel, et al. 2011; Kofler, Pandey, et al. 2011; Raineri et al. 2012; Lynch et al. 2014; Nelson and Hughes 2015). This allows unprecedented insight into the molecular evolution of virus populations, both within their hosts and during transmission, which are otherwise difficult to characterize. In particular, the software SNPGenie (Chapter 3; Nelson et al. 2015; https://github.com/hugheslab/snpgenie) can be used to estimate population genetic parameters from pooled-sequencing variant data, including π and gene diversity at nonsynonymous and synonymous sites, at a range of levels including single nucleotides, codons, genes, and genomes (populations).

Bailey et al. (2014) showed that overall $\pi_N < \pi_S$ for krc1 and krc2, suggesting that purifying selection dominates their evolution, as is observed for virtually all viruses studied to date (Holmes 2009). They further demonstrated that π_N tends to be higher in 3'-proximal ORFs, with ORFs 3 and 5 containing regions in which $\pi_N > \pi_S$, suggestive of overdominant selection for immune escape. Finally, they noted that a positive correlation exists between viremia (viral load) and both π_N and π_S, with all three measures being significantly higher in krc1 than krc2. Nelson and Hughes (2015) further noted that viremia should be directly related to effective population size (N_e), suggesting that within-host viral data can be used to elucidate the relative contributions of N_e, mutation rates, and generation times in viral evolution.

In the present study, we use recent advances in SNPGenie to develop approaches for answering these and other questions about the evolution of viruses. Using the pooled-sequencing data of Bailey et al. (2014), SNPs were called relative to *de novo* reference sequences, and quality filtering was used to eliminate SNPs with Q < 25, reads < 100 nt in length, and estimated frequencies < 5%. Because these viruses share a unique genome in which all ORFs overlap at their termini, with one ORF being entirely subsumed by another (TF within 1a), we first estimated π for all sites, and then separately for those which do and do not overlap multiple reading frames. The results allowed us to compare relative constraint of nonsynonymous and synonymous sites in both types of regions. Surprisingly, unlike in other viruses such as papillomaviruses (Hughes and Hughes 2005), sites in krc1 and krc2 which overlap multiple reading frames tend to be less constrained in terms of nonsynonymous changes than non-overlapping sites. Our results suggest that this is due to an enrichment of viral epitopes in these regions, and that

purifying selection still acts by limiting the specific nonsynonymous changes that can occur in overlapping regions.

As recent evidence has shown that selection acting on viral epitopes can often not be detected at the whole-gene levels (Bailey et al., *in press*), we next undertake unsupervised nonsynonymous peak discovery using a sliding window approach to detect candidate epitope regions. Several candidate epitopes are identified, corroborating earlier findings that suggest important regions exist which cannot be identified using either single-codon or whole-gene statistics. Peaks in both viruses are remarkably concordant in terms of ORFs and positions within ORFs, suggesting that viable escape mutations are somewhat limited. Insertions and deletions (indels) also seem to play a key role in immune escape, as the ORFs exhibiting the most indels also contained nonsynonymous peaks.

Elucidating the population genetic factors contributing to viral evolution requires not only an estimate of N_e, but also an estimate of the viral mutation rate. To this end, we present a simple population genetic estimator based on within- and between-population synonymous genetic distance for use with longitudinal population data. One co-infected RC host was sampled at two time points, allowing us to make such estimates for both krc1 and krc2, which to our knowledge are the first of their kind. Having obtained estimated mutation rates, we then explore the relative contributions of mutation and selection in the evolution of these viruses. Our results suggest that the majority of nonsynonymous evolutionary change is driven by mutation pressure, but also that most deleterious mutations probably persist at very low frequencies in viral populations.

Finally, we estimate gene diversity (*H*) separately for nonsynonymous and synonymous sites, as well as the Tajima's (*D*) statistic, showing that the quality control filtering currently necessary for pooled-sequencing SNP calling severely limit the utility of these measures.

4.2 Results

4.2.1 Evolutionary Constraint in Non-overlapping and Overlapping ORF Regions

Previous work with SHFV viruses krc1 and krc2 has shown that $\pi_N < \pi_S$ for both, evidencing widespread purifying selection and demonstrating the relative constraint of nonsynonymous sites as compared to synonymous sites (Bailey et al. 2014). However, many viral genomes exhibit higher sequence constraint in regions which overlap multiple ORFs as opposed to those which do not overlap (Belshaw et al. 2007; Sabath 2009). This is presumably because mutations in these regions are likely to alter multiple protein products, increasing the probability that they will have deleterious effects (*e.g.*, papillomaviruses; Hughes and Hughes 2005). However, some viruses contain important viral epitopes within such overlapping regions, the result being rapid sequence change driven by overdominant selection (*e.g.*, simian immunodeficiency virus; Hughes et al. 2001). Previous analyses with krc1 and krc2 found that π_N peaks occurring in overlapping regions specific to ORFs 3 and 5 correspond to π_S peaks in the overlapping ORF, suggesting the possibility that positive selection in one ORF may be accompanied by purifying selection in the overlapping ORF at the same genomic positions (Bailey et al. 2014).

To compare evolutionary constraint at coding residues which do and do not overlap multiple ORFs, we analyzed regions of each type separately for each virus. Because co-infection with both krc1 and krc2 does not impact the nucleotide diversity of either virus (Bailey et al. 2014), we included both mono- and co-infections in our analyses. The highest proportions of overlap occurred in ORF TF (overlaps 1a), with 100% overlap in both viruses, and in ORF 2a, with 88.89% and 86.99% overlap in krc1 and krc2, respectively. ORF 1b was the second largest ORF and exhibited the least proportion of overlap (overlaps 1a and 2a'), with 1.55% and 2.80% overlap for krc1 and krc2, respectively (Table 4.1). There was no significant tendency for ORFs to be longer in either virus. However, differences in length were significant for all ORFs ($\alpha = 0.05$ with Bonferroni correction for 14 Wilcoxon Signed Rank tests). Of these differences, the greatest differences in total non-STOP codons were in ORFs 4' (28 more codons in krc1, 26 of which were non-overlapping), 1b (21 more codons in krc2, 19 of which were overlapping), and 3' (18 more codons in krc1, 6 of which were overlapping). There were also instances in which the amount of overlap shifted greatly but total ORF length was relatively preserved. The most dramatic examples were ORF 3 (45 more overlapping codons but 43 fewer non-overlapping codons in krc1), 2b (26 more overlapping codons but 25 fewer non-overlapping codons in krc1), and 4 (18 more overlapping codons but 17 fewer non-overlapping codons in krc1).

For krc1, mean $\pi_N = 0.00264$ (± 0.00021 S.E.M.) and mean $\pi_S = 0.01529$ (± 0.00142). The mean values of π_N for non-overlapping (NOL) and overlapping (OL) residues were $\pi_{N\text{-}NOL} = 0.00183$ (± 0.00015) and $\pi_{N\text{-}OL} = 0.00585$ (± 0.00049),

Table 4.1. Mean number of overlapping (% of total) and total codons by ORF.

	krc1		krc2	
ORF	**Overlapping**	**Total**	**Overlapping**	**Total**
1a	222.38 (10.85%)	2049	223.33 (10.88%)	2053
TF	220 (100%)	220	221 (100%)	221
1b	22.71 (1.55%)	1465.04	41.67 (2.80%)	1486
2a’	69.67 (29.90%)	233	82.67 (36.91%)	224
3’	74.67 (33.79%)	221	68.67 (33.83%)	203
4’	29.67 (15.22%)	195	28 (16.77%)	167
2a	72 (88.89%)	81	71.33 (86.99%)	82
2b	135 (65.53%)	206	108.93 (53.16%)	204.93
3	117.88 (59.15%)	199.30	72.59 (36.92%)	196.63
4	50.88 (29.37%)	173.22	32.67 (18.99%)	172
5a	49.74 (74.53%)	66.74	42 (71.19%)	59
5	51.74 (20.93%)	247.22	44 (18.69%)	235.37
6	11 (6.79%)	162	9.22 (5.75%)	160.22
7	8.67 (7.23%)	119.96	6.89 (6.19%)	111.22

ORFs descend in the table from 5’ to 3’ by start site. Mean numbers of codons were determined by averaging the length of the respective ORF’s consensus sequences across all within-host populations analyzed, 23 isolates for krc1 and 27 isolates for krc2.

respectively. The mean values of π_S for non-overlapping and overlapping residues were $\pi_{S\text{-}NOL}$ = 0.01616 (± 0.00164) and $\pi_{S\text{-}OL}$ = 0.01186 (± 0.00108), respectively. For krc2, π_N = 0.00210 (± 0.00019) and π_S = 0.00905 (± 0.00096); π_N for non-overlapping and overlapping residues were $\pi_{N\text{-}NOL}$ = 0.00180 (± 0.00019) and $\pi_{N\text{-}OL}$ = 0.00322 (± 0.00034); and π_S for non-overlapping and overlapping residues were $\pi_{S\text{-}NOL}$ = 0.00959 (± 0.00105) and $\pi_{S\text{-}OL}$ = 0.00689 (± 0.00074). All values of π were higher in krc1 than in

krc2. This difference was significant for $\pi_{N\text{-}OL}$ ($P < 0.01$), π_S ($P = 0.0008$), $\pi_{S\text{-}NOL}$ ($P = 0.002$), and $\pi_{S\text{-}OL}$ ($P = 0.0005$), but not significant for π_N or $\pi_{N\text{-}NOL}$ (two-sample T-tests).

At the genome level, $\pi_N < \pi_S$ was significant for both viruses, including within both overlapping and non-overlapping regions ($P < 0.0001$; paired T-tests), constituting strong evidence of purifying selection that is consistent with previous results (Bailey et al. 2014). However, contrary to expectation, $\pi_{N\text{-}OL}$ significantly exceeded $\pi_{N\text{-}NOL}$ in krc1 ($P < 0.0001$) and in krc2 ($P = 0.0003$; paired T-tests), implying decreased constraint for nonsynonymous changes at overlapping sites. The opposite pattern was observed for π_S, with $\pi_{S\text{-}NOL}$ significantly exceeding $\pi_{S\text{-}OL}$ in both krc1 and krc2 ($P = 0.011$ and $P = 0.0006$, respectively; paired T-tests).

Even when overlapping ORFs contain known epitopes (*e.g.*, the overlap between simian immunodeficiency virus ORFs tat and vpr; Hughes et al. 2001), a negative correlation can be observed between the proportion of an ORF's overlap and its π_S. This is because synonymous changes in one ORF are likely to be nonsynonymous in the alternative ORF, making them subject to purifying selection that will lower π_S. Between ORFs, the distributions of π estimates for both viruses were positively skewed and departed significantly from normality ($P < 0.001$; Shapiro-Wilk test), necessitating the use of nonparametric tests. In keeping with our surprising result that overlapping regions had a higher π_N (*i.e.*, $\pi_{N\text{-}OL} > \pi_{N\text{-}NOL}$), there was no significant correlation between an ORF's π_S and its proportion of overlap. Thus, differences among ORFs with respect to proportions of overlap do not explain the variation in π_S observed in either virus.

We next analyzed each ORF separately to explore differences between

(A)

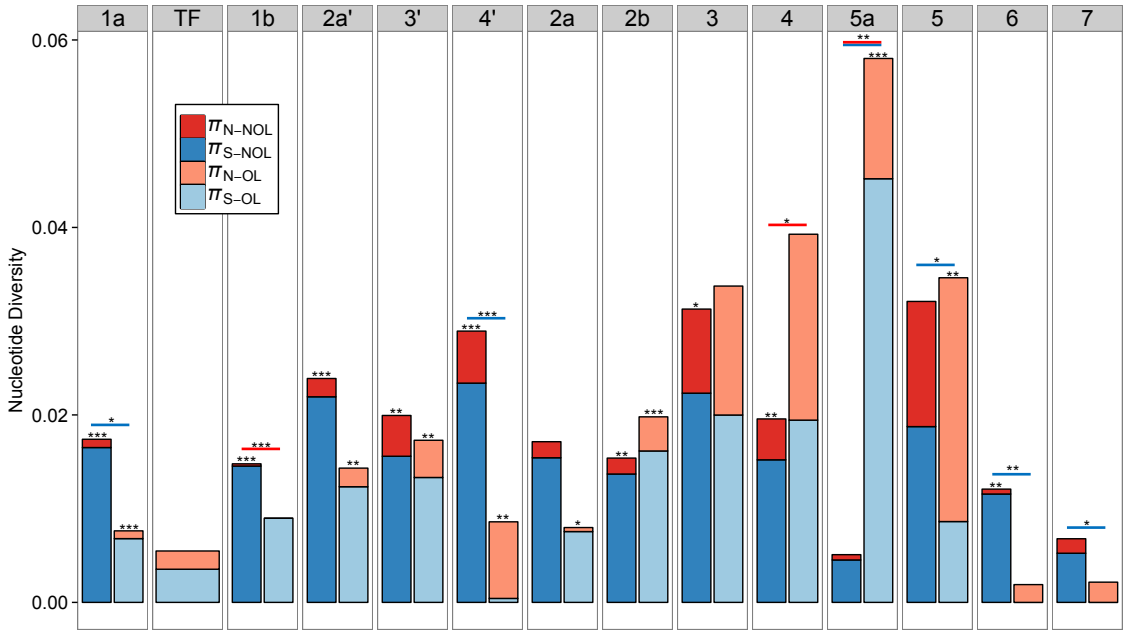
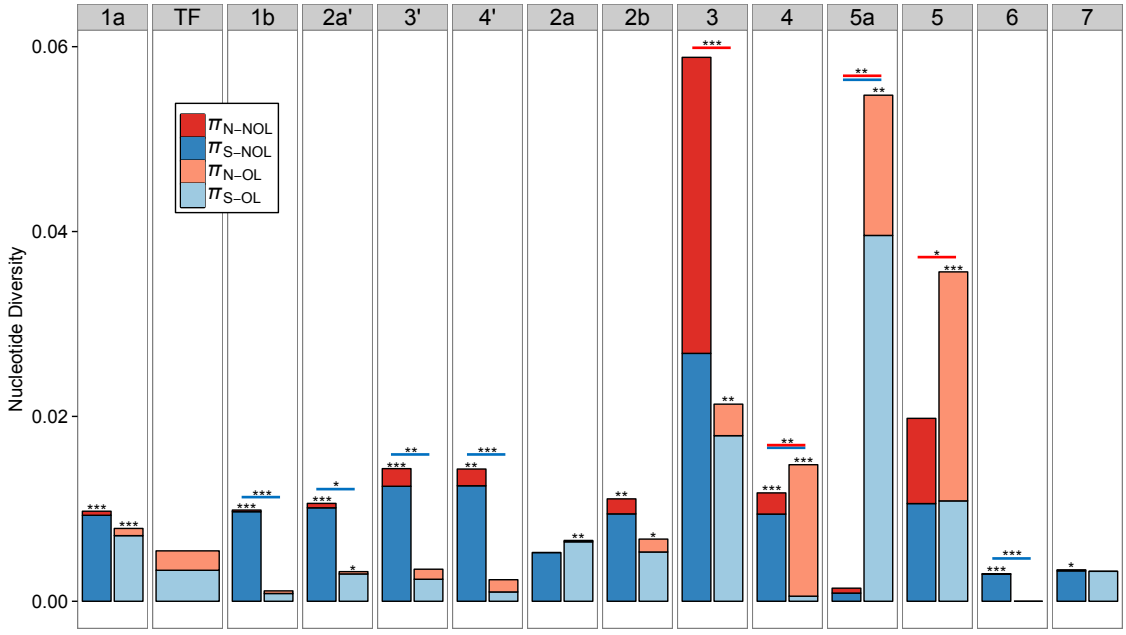


Figure 4.2. Nonsynonymous and synonymous nucleotide diversity in overlapping (OL) and non-overlapping (NOL) ORF sites. Results shown separately for krc1 (A) and krc2 (B). ORFs are ordered 5' to 3', but panel width is not proportional to ORF size. Colors are red for π_N and blue for π_S. Within bar pairs for each ORF, darker colors (left bars) represent non-overlapping sites and lighter colors (right bars) represent overlapping sites. ORF TF contains only overlapping residues. Horizontal red lines indicate a significant difference between an ORF's $\pi_{N\text{-}NOL}$ and $\pi_{N\text{-}OL}$; horizontal blue lines indicate a significant difference between an ORF's $\pi_{S\text{-}NOL}$ and $\pi_{S\text{-}OL}$. Where there are significant differences between π_N and π_S within overlapping ($\pi_{N\text{-}OL}$ vs. $\pi_{S\text{-}OL}$) or within non-overlapping ($\pi_{N\text{-}NOL}$ vs. $\pi_{S\text{-}NOL}$) residues, these differences are indicated above each bar. All tests were Wilcoxon Signed Rank tests, with significance levels determined using a Bonferroni correction for 14 tests: * for $\alpha < 0.05$; ** for $\alpha < 0.01$; and *** for $\alpha < 0.001$.

overlapping regions in different parts of the viral genomes. In krc1's non-overlapping sites, $\pi_{N\text{-}NOL} < \pi_{S\text{-}NOL}$ occurred in all ORFs, being significant in all except 2a, 5a, 5, and 7 ($\alpha = 0.05$ with Bonferroni correction for 14 Wilcoxon Signed Rank tests used throughout). On the other hand, in its overlapping sites, $\pi_{N\text{-}OL} < \pi_{S\text{-}OL}$ was significant only

(B)



... see previous page.

---

in ORFs 1a, 2a', 3', 2a, 2b, and 5a, while the opposite pattern of $\pi_{N\text{-}OL} > \pi_{S\text{-}OL}$ was significant in ORFs 4' and 5, and also present but not significant in ORF 4 (Figure 4.2A). Similarly, in krc2's non-overlapping sites, $\pi_{N\text{-}NOL} < \pi_{S\text{-}NOL}$ was significant in all ORFs except 2a, 3, 5a, and 5, while the opposite pattern of $\pi_{N\text{-}NOL} > \pi_{S\text{-}NOL}$ occurred in ORF 3 but was not significant. Also similar to krc1, in krc2's overlapping sites, $\pi_{N\text{-}OL} < \pi_{S\text{-}OL}$ in krc2 was significant only in ORFs 1a, 2a', 2a, 2b, 3, and 5a, while the opposite pattern of $\pi_{N\text{-}OL} > \pi_{S\text{-}OL}$ was highly significant in ORFs 4 and 5, and present but not significant in ORF 4' (Figure 4.2B).

It is particularly noteworthy that several ORFs saw a reversal of their π_N/π_S ratio between residues which do and do not overlap multiple ORFs. In krc1, a reversal from $\pi_{N\text{-}NOL} < \pi_{S\text{-}NOL}$ to $\pi_{N\text{-}OL} > \pi_{S\text{-}OL}$ was significant in ORF 4' ($P < 0.01$; Wilcoxon Signed

Rank test), and present but not significant in ORFs 4 and 5. In krc2, the same reversal was significant in ORF 4 ($P < 0.001$; Wilcoxon Signed Rank test), and present but not significant in ORFs 4' and 5. The opposite reversal from $\pi_{N\text{-}NOL} > \pi_{S\text{-}NOL}$ to $\pi_{N\text{-}OL} < \pi_{S\text{-}OL}$ was present but not significant in ORF 3. Thus ORFs 4', 4, and 5 all exhibit strong evidence of purifying selection in their non-overlapping residues while concurrently exhibiting evidence of overdominant selection in their overlapping residues.

In summary, π_S significantly exceeded π_N in both overlapping and non-overlapping ORF regions of both viruses, evidencing widespread purifying selection. However, contrary to what might have been expected based on functional constraint, we observed that $\pi_{N\text{-}OL}$ is significantly higher than $\pi_{N\text{-}NOL}$ in both viruses. On the other hand, $\pi_{S\text{-}NOL}$ was significantly larger than $\pi_{S\text{-}OL}$ for both viruses. Two ORFs—ORF 4' in krc1 and ORF 4 in krc2—showed a particularly interesting pattern in which strong purifying selection in non-overlapping regions was accompanied by a strong signature of overdominant positive selection in overlapping regions.

4.2.2 Unsupervised Epitope Discovery: Identifying Nonsynonymous Peaks Using Sliding Windows of $\pi_N > \pi_S$

One possible explanation for the lack of consistent heightened nonsynonymous constraint in overlapping regions is that these loci may be particularly enriched in epitopes undergoing overdominant selection for immune escape. It was thus necessary to identify putative epitopes. Unfortunately, a statistical signal often cannot be detected at the ORF level. For example, recent work with African Green Monkey viruses demonstrates that the signal of $\pi_N < \pi_S$ can be lost within ORFs containing known epitopes that are

otherwise constrained by purifying selection, presumably because localized overdominant selection cancels out the background signal of purifying selection at low genomic resolution (Bailey et al., *in press*). For this reason, it is necessary to perform sliding window analyses at a biologically meaningful scale. Such regions are characterized by $\pi_N > \pi_S$, even when this does not hold for the entire ORF (Hughes and Nei 1988; Halliburton 2004). The average size of antigenic peptide fragments presented by host MHC class I receptors to CD8+ (cytotoxic) T-cells is 9 amino acids (Rammensee et al. 1995; Evans et al. 1999; Hughes et al. 2001). Thus, in terms of host-pathogen coevolution, we might expect to observe peaks in nonsynonymous polymorphism in 9-codon windows as a result of overdominant selection.

Unsupervised nonsynonymous peak discovery was performed using 9-codon sliding windows of π_N/π_S across the krc1 and krc2 genomes, including aligned data from all isolates of each. For ORFs that differed in length between or within viruses, products were translated, aligned at the amino acid level, and this alignment was then imposed on the DNA sequence. We defined nonsynonymous peaks as windows in which π_N exceeded not only that window's π_S, but also the overall π_S of the respective ORF, which was always greater than 0. This was necessary to preclude the possibility of identifying peaks that were due either to mutational hotspots or else to stochastic (low) fluctuations in π_S, as might be imposed by codon bias. Overlapping peaks were concatenated and the resultant regions were end-trimmed to remove codons lacking polymorphism.

Our approach identified 20 nonsynonymous peaks, 12 peaks in krc1 with median length 15 (± 9.5 I.Q.R.) and 8 peaks in krc2 with median length 26.5 (± 15) (Table 4.2).

Table 4.2. Peaks of nonsynonymous viral polymorphism suggestive of overdominant selection and epitope function.

	ORF	ORF π_N	ORF π_S	Peak	Peak π_N	Peak π_S	Start	Stop	Codons	Isolates with nonsyn. SNPs
krc1	TF	0.00195	0.00354	1◦	0.00899	0.00108	162	172	11	47.8%
				2◦	0.00422	0.00000	194	207	14	34.8%
				3◦	0.00832	0.00000	215	218	4	21.7%
	3’	0.00422	0.01482	1◦	0.01568	0.00000	16	31	16	65.2%
				2	0.01905	0.00801	178	184	7	47.8%
	4’	0.00592	0.01978	1◦	0.01988	0.00000	16	24	9	78.3%
				2	0.04978	0.04004	132	144	13	87.0%
	3	0.01180	0.02093	1◦	0.03392	0.00566	46	63	18	91.3%
				2*◦	0.03184	0.02318	152	168	17	91.3%
	4	0.00892	0.01644	1◦	0.02914	0.00183	23	48	26	78.3%
	5	0.00971	0.03531	1◦	0.04680	0.00892	15	42	28	91.3%
				2*	0.08344	0.05206	72	100	29	95.7%
krc2	TF	0.00210	0.00335	1◦	0.00839	0.00426	90	123	34	70.4%
				2◦	0.00441	0.00000	170	183	14	44.4%
	4’	0.00174	0.01041	1	0.01680	0.00875	69	79	11	51.9%
	3	0.02179	0.02330	1*	0.02679	0.02112	120	132	13	96.3%
				2*	0.10550	0.05369	136	163	28	100%
	4	0.00454	0.00778	1◦	0.01752	0.00522	11	41	31	85.2%
	5	0.01082	0.03005	1◦	0.03459	0.00921	12	38	27	85.2%
				2*	0.06393	0.04511	64	89	26	96.3%

Peaks were identified conservatively as 9-codon sliding windows in which π_N exceeded 0, the respective window’s π_S, and the mean value of π_S for the ORF. ORF π_N and π_S values include all sites in the ORF. Windows were combined if they overlapped and end-trimmed to remove codons with no within-host nonsynonymous polymorphism. Start and stop sites refer to the median ORF codon coordinates in between-isolate sequence alignments (*i.e.*, site 1 is codon 1, which usually overlaps the previous ORF). *Peak contains indels between isolates. ◦Peak is present in overlapping regions of the ORF.

Of the 14 ORFs examined, 5 contained peaks in both viruses: TF (three peaks in krc1 and two peaks in krc2), 4’ (two peaks in krc1 and one peak in krc2), 3 (two peaks in both viruses), 4 (one peak in both viruses), and 5 (one peak in both viruses). ORF 3’ contained

two peaks in krc1 but none in krc2. To this point, it is significant that ORF 3' experienced an 18-codon deletion in krc2 as compared to krc1 (Table 4.1). Alignments between the viruses show that this was primarily achieved via two distinct 9-codon deletions starting at codons 107 and 158 of ORF 3' in krc1, suggesting deletion of two epitopes. However, alignments indicate that neither of these deletions is located in residues aligning to either nonsynonymous peak of ORF 3' in krc1. The remaining 8 ORFs contained no peaks in either virus.

There was significant concordance between the two viruses for the number of peaks in each ORF (P = 0.0030; Fisher's Exact test), as well as for their genomic locations, with 10 of 20 peaks (50%) overlapping the same locations in both viruses. For ORF TF, the last 4 codons of peak 1 in krc1 overlapped peak 2 in krc2 (32.0% concordance). For ORF 4', the first 9 codons of peak 2 in krc1 overlapped a deletion in krc2 (69.2% overlap). For ORF 3, the first 15 codons of peak 2 in krc1 overlapped with the last 15 codons of peak 2 in krc2 (61.2% concordance). For ORF 4, the first 24 codons of the peak in krc1 overlapped the last codons of the peak in krc2 (84.2% concordance). ORF 5 had the most concordant peaks between viruses; peak 1 of krc1 was contained entirely within peak 1 of krc2 (98.2% concordance), which itself contained a 4-codon deletion, while the last 27 codons of peak 2 of krc1 overlapped the first codons of peak 2 in krc2 (91.5% concordance). Here, ORF 5 peak 2 in krc1 contained a central 2-codon deletion, while the same peak in krc2 commonly contained 6- and 3-codon deletions.

Within-host populations of the same virus also exhibited indels within several peaks (Table 4.3). In krc1 ORF 3 peak 2, 14 of 21 (66.7%) isolates exhibiting peak

Table 4.3. Isolates exhibiting peaks of nonsynonymous viral polymorphism.

Virus	ORF	Peak	Isolates with non-silent polymorphism
krc1	TF	1◦	RC05, RC06, RC07, RC09, RC10, RC18, RC22, RC25, RC31, RC34, RC51
		2◦	RC05, RC09, RC18, RC25, RC33, RC34, RC45, RC61
		3◦	RC09, RC25, RC30, RC34, RC61
	3’	1◦	RC05, RC06, RC08, RC10, RC18, RC30, RC31, RC33, RC34, RC40, RC44, RC45, RC56, RC60, RC61
		2	RC09, RC10, RC30, RC33, RC34, RC40, RC44, RC45, RC51, RC60, RC61
	4’	1◦	RC05, RC06, RC08, RC09, RC10, RC18, RC22, RC25, RC30, RC31, RC33, RC34, RC40, RC45, RC54, RC56, RC60, RC61
		2	RC05, RC06, RC07, RC08, RC10, RC13, RC22, RC25, RC30, RC31, RC33, RC34, RC40, RC44, RC45, RC51, RC54, RC56, RC60, RC61
	3	1◦	RC05, RC06, RC07, RC08, RC09, RC10, RC13, RC18, RC22, RC25, RC30, RC31, RC33, RC34, RC40, RC44, RC45, RC54, RC56, RC60, RC61
		2*◦	RC05^{+}, RC06^{+}, RC07°, RC08^{+}, RC09, RC13°, RC18, RC22^{+}, RC25^{+}, RC28^{+}, RC30, RC31^{+}, RC33^{+}, RC34^{+}, RC40^{+}, RC44^{+}, RC45^{+}, RC51, RC54^{+}, RC56^{+}, RC60
	4	1◦	RC05, RC06, RC07, RC08, RC10, RC13, RC18, RC22, RC25, RC30, RC31, RC33, RC34, RC40, RC44, RC51, RC56, RC60
	5	1◦	RC05, RC06, RC07, RC08, RC09, RC10, RC13, RC18, RC22, RC25, RC30, RC33, RC34, RC40, RC44, RC45, RC51, RC54, RC56, RC60, RC61
		2*	RC05^{+}, RC06^{+}, RC07^{+}, RC08^{+}, RC09^{+}, RC10^{+}, RC13^{+}, RC22^{+}, RC25^{+}, RC28^{+}, RC30^{+}, RC31^{+}, RC33^{+}, RC34^{+}, RC40^{+}, RC44^{+}, RC45^{+}, RC51^{+}, RC54^{+}, RC56^{+}, RC60^{+}, RC61^{+}
krc2	TF	1◦	RC06, RC08, RC10, RC14, RC15, RC18, RC20, RC22, RC25, RC31, RC33, RC34, RC39, RC42, RC51, RC54, RC55, RC56, RC60
		2◦	RC06, RC08, RC18, RC20, RC22, RC28, RC31, RC33, RC34, RC39, RC40, RC56
	4’	1	RC06, RC07, RC10, RC14, RC15, RC18, RC26, RC31, RC33, RC34, RC39, RC40, RC42, RC61
	3	1*	RC06^{+}, RC07°, RC08^{+}, RC10^{+}, RC13°, RC14^{+}, RC15^{+}, RC18^{+}, RC20^{+}, RC22^{+}, RC25^{+}, RC26^{+}, RC28^{+}, RC31^{+}, RC33^{+}, RC34^{+}, RC39^{+}, RC40^{+}, RC42^{+}, RC44°, RC51^{+}, RC54^{+}, RC55^{+}, RC56, RC60°, RC61^{+}
		2*	RC05^{+}, RC06^{+}, RC07^{+}, RC08, RC10^{+}, RC13^{+}, RC14^{+}, RC15^{+}, RC18^{+}, RC20^{+}, RC22^{+}, RC25^{+}, RC26^{+}, RC28^{+}, RC31^{+}, RC33^{+}, RC34^{+}, RC39^{+}, RC40^{+}, RC42^{+}, RC44^{+}, RC51^{+}, RC54°, RC55^{+}, RC56^{+}, RC60^{+}, RC61^{+}
	4	1◦	RC05, RC06, RC08, RC10, RC14, RC15, RC18, RC20, RC22, RC25, RC26, RC28, RC31, RC33, RC34, RC40, RC42, RC44, RC51, RC55, RC56, RC60, RC61
	5	1◦	RC05, RC06, RC08, RC10, RC13, RC14, RC15, RC18, RC20, RC22, RC25, RC26, RC28, RC31, RC33, RC34, RC39, RC42, RC51, RC55, RC56, RC60, RC61
		2*	RC05°, RC06°, RC07°, RC08^{+}, RC10^{+}, RC14^{+}, RC15^{+}, RC18^{+}, RC20^{+}, RC22^{+}, RC25°, RC26^{+}, RC28^{+}, RC31^{+}, RC33^{+}, RC34°, RC39^{+}, RC40^{+}, RC42^{+}, RC44°, RC51^{+}, RC54°, RC55°, RC56°, RC60^{+}, RC61^{+}

Peaks (Table 4.2) from red colobus (RC) hosts: *Peak contains indels between isolates. ◦Peak is present in overlapping regions of the ORF. $^{+}$Isolate contains both deletions and nonsynonymous SNPs. °Isolate contains deletions but no nonsynonymous SNPs.

polymorphism contained both alignment gaps and nonsynonymous polymorphism, while 2 more contained gaps only. These gaps often occurred at the fifth codon of the peak. In krc1 ORF 5 peak 2, all isolates exhibiting peak polymorphism contained both gaps and nonsynonymous polymorphism. These gaps tended to be central in the peak, involving a median of 9 codons. Interestingly, the one isolate (RC18) which did not contain nonsynonymous polymorphism within this peak region was also the only isolate which had no gaps. In krc2 ORF 3 peak 1, 21 of 26 (80.8%) isolates exhibiting peak polymorphism contained both alignment gaps and nonsynonymous polymorphism, while 4 more contained gaps only. Further, similar to krc1 ORF 3 peak 2, this peak contained a gap at the fifth codon in all but one isolate (RC05). Thus, there is a strong resemblance between ORF 3 peak 2 in krc1 and ORF 3 peak 1 in krc2. In krc2 ORF 3 peak 2, 25 of 27 (92.6%) isolates exhibiting peak polymorphism contained both alignment gaps and nonsynonymous polymorphism, while 1 more contained gaps only. These gaps tended to be central in the peak, also involving a median of 9 codons. Finally, in krc2 ORF 5 peak 2, 17 of 26 (63.4%) isolates exhibiting peak polymorphism contained both alignment gaps and nonsynonymous polymorphism, while 9 more contained gaps only. These gaps tended to occur in the first half of the peak, involving a median of 5 codons.

Although only 20.15% of all genomic positions in krc1 overlapped multiple ORFs, 75% of nonsynonymous peaks were in overlapping regions. Likewise, although only 18.89% of all genomic positions in krc2 overlapped multiple ORFs, 50% of nonsynonymous peaks were in overlapping regions (Table 4.2). Thus, overlapping regions were enriched in nonsynonymous peaks as compared to the random expectation. In all instances, nonsynonymous peaks were only present in one of the two overlapping

ORFs, *i.e.*, nonsynonymous peaks located in two overlapping ORFs never shared any genomic positions. For example, in both viruses, ORF TF is entirely subsumed by ORF 1a, but TF contains multiple nonsynonymous peaks while 1a has none.

The preponderances of nonsynonymous peaks in overlapping regions may explain our unexpected result that $\pi_{N\text{-}OL} > \pi_{N\text{-}NOL}$ in krc1 ORF 4 and krc2 ORFs 4 and 5. To see whether this was due primarily to the nonsynonymous peaks we identified, we re-calculated $\pi_{N\text{-}OL}$ and $\pi_{N\text{-}NOL}$ for these ORFs with nonsynonymous peaks excluded. For krc1 ORF 4, $\pi_{N\text{-}OL}$ dropped from 0.0198 to 0.0133, but this still exceeded $\pi_{N\text{-}NOL}$ = 0.00436. On the other hand, the pattern dramatically reversed in krc2 ORF 4, with $\pi_{N\text{-}OL}$ dropping from 0.0142 to 0.000480, far below the non-peak $\pi_{N\text{-}NOL}$ = 0.00163. For krc2 ORF 5, $\pi_{N\text{-}OL}$ dropped markedly from 0.0248 to 0.00339, but still exceeded the non-peak $\pi_{N\text{-}NOL}$ = 0.000639. Thus, the unexpected pattern of $\pi_{N\text{-}OL} > \pi_{N\text{-}NOL}$ can sometimes but not always be explained by the nonsynonymous peak regions we identified.

4.2.3 Effects of Nonsynonymous SNPs on Overlapping ORFs Within Nonsynonymous Peaks

The majority of nonsynonymous SNPs occurring in one ORF will also be nonsynonymous in an overlapping ORF. The exact proportion depends on the specific codons used and whether the reading frames are offset by 1 or 2 positions. Because nonsynonymous peaks from overlapping ORFs never fall over the same genomic positions in our data, we hypothesized that the residues exhibiting nonsynonymous polymorphism in one ORF would be relatively constrained in the alternative overlapping ORF. Based on this hypothesis, we predicted that nonsynonymous changes in the peak-

containing ORF would occur disproportionately so as to cause synonymous changes in the alternative ORF. This would constitute a test for purifying selection that controls for overdominant selection in one frame.

Across all overlapping ORF regions containing nonsynonymous peaks in krc1, 64.30% of all possible nonsynonymous changes also resulted in a nonsynonymous change in the overlapping ORF; however, only 32.06% of the observed SNPs did so. Likewise, across all overlapping ORF regions containing nonsynonymous peaks in krc2, 63.49% of all possible nonsynonymous changes also resulted in a nonsynonymous change in the overlapping ORF; however, only 29.82% of the observed SNPs did so. Thus, nonsynonymous changes in the peak-containing ORF indeed occurred disproportionately so as to cause synonymous changes in the alternative ORF. This suggests that purifying selection acting on one ORF can constrain the nonsynonymous changes that are accepted in an overlapping ORF.

This pattern is even more illuminating when viewed separately for each ORF (Table 4.4). For each overlapping region, nonsynonymous changes were analyzed in both the peak and in the non-peak (remainder) residues. For 16 peak and remainder regions in krc1, nonsynonymous changes in the peak-containing ORF caused fewer than expected nonsynonymous changes in the overlapping ORF in all but one instance, and this difference was significant in 13 of the regions ($\alpha = 0.05$ with Bonferroni correction for 23 Exact Binomial tests). Likewise, for the 7 peak and remainder regions in krc2, nonsynonymous changes in the peak-containing ORF caused fewer than expected nonsynonymous changes in the overlapping ORF in all cases, and this difference was

Table 4.4. Effects of nonsynonymous SNPs on overlapping ORFs.

Virus	ORF	Alt. (OL) ORF	Non-syn. peak	No. cod-ons	Prop. OL region	Non-syn. SNPs	Prop. exp. nonsyn. in alt. ORF	Prop. obs. nonsyn. in alt. ORF	*P*-value
krc1	TF	1a	1	12	5.45%	15	61.12%	6.67%	< 0.0001***
			2	15	6.82%	9	67.68%	0%	< 0.0001***
			3	5	2.27%	5	65.34%	0%	0.005
			rem.	188	85.45%	41	65.91%	21.95%	< 0.0001***
	3'	2a'	1	17	34.00%	45	58.43%	11.11%	< 0.0001***
			rem.	33	66.00%	7	59.59%	0.00%	0.0018*
	4'	3'	1	10	38.46%	29	63.64%	3.45%	< 0.0001***
			rem.	16	61.54%	8	63.01%	0.00%	0.0004**
	3	2b	1	18	26.47%	106	62.96%	30.19%	< 0.0001***
			rem.	50	73.53%	42	63.27%	19.05%	< 0.0001***
		4	2	19	36.54%	102	67.77%	38.24%	< 0.0001***
			rem.◦	33	63.46%	31	63.55%	83.87%	0.0234
	4	3	1	27	51.92%	120	57.28%	19.17%	< 0.0001***
			rem.◦	25	48.08%	70	64.73%	60.00%	0.4531
	5	5a	1	28	58.33%	202	66.67%	43.07%	< 0.0001***
			rem.	20	41.67%	10	61.16%	10.00%	0.0013*
krc2	TF	1a	1	35	15.84%	51	61.53%	9.80%	< 0.0001***
			2	15	6.79%	18	64.33%	0.00%	< 0.0001***
			rem.	171	77.38%	38	64.20%	13.16%	< 0.0001***
	4	3	1	23	69.70%	87	52.08%	12.64%	< 0.0001***
			rem.	10	30.30%	2	65.52%	0.00%	0.1189
	5	5a	1	28	66.67%	173	69.60%	52.60%	< 0.0001***
			rem.	14	33.33%	10	63.83%	10.00%	0.0007*

Peak numbers refer to those in Table 4.2. "Prop. OL region" is the proportion of the entire overlapping region between two ORFs that is occupied by the feature in question. No peaks in different ORFs overlapped one another, but two did co-exist side by side in the same region of overlap between krc1 ORFs 3 and 4. The proportion of alternative ORF SNPs expected to be nonsynonymous was calculated as the fraction of all possible nonsynonymous SNPs in the peak ORF which were also nonsynonymous in the alternative (overlapping; OL) ORF, corresponding to p_0 in Exact Binomial tests. The symbol ◦ indicates overlapping but non-peak remainder residues (rem.) that overlap nonsynonymous peaks in an overlapping ORF. *P*-values refer to the results of Exact Binomial tests, with significance levels determined using a Bonferroni correction for 23 tests: * for $\alpha < 0.05$; ** for $\alpha < 0.01$; and *** for $\alpha < 0.001$.

significant in 6 ($\alpha = 0.05$ with Bonferroni correction for 23 Exact Binomial tests). This difference was more pronounced in all remainder regions, excepting those of ORF TF in both viruses and the ORF 3/ORF 4 overlapping region of krc1.

One exception occurred in krc1, namely, 83.87% of nonsynonymous SNPs in the remainder of ORF 3 peak 2 were also nonsynonymous in the overlapping ORF 4. Interestingly, this remainder region is occupied by a distinct nonsynonymous peak in the overlapping ORF 4 (peak 1). Complementarily, 60.00% of nonsynonymous SNPs in the remainder of ORF 4 peak 1 were also nonsynonymous in the overlapping ORF 3 peak 2. This was the only instance in which two peaks overlapped one another's remainder regions in either viral genome, and explains the higher proportion of nonsynonymous overlapping changes as compared to other overlapping regions.

4.2.4 Viremia and the Strength of Selection

The differences between π_N and π_S within krc1 and krc2 provide a measure of the relative strength of purifying selection in the two viruses. According to the neutral theory, the efficacy of selection is directly proportional to the effective population size, N_e, and mutations having fitness effects of a magnitude much less than $1/N_e$ will behave essentially as if neutral (Wright 1931; Kimura 1983; Lynch 2007a). Thus, as N_e increases, the range of fitness effects dominated by random genetic drift shrinks.

In terms of population genetics, viremia may be considered a proxy for N_e. Bailey et al. (2014) have shown that viremia is significantly higher in krc1 than in krc2, implying that krc1 has a larger N_e. For all co-infected hosts in our study, viremia

Table 4.5. Red colobus (RC) host viremia measures for krc1 and krc2.

Isolate	krc1	krc2
RC05	9.20×10^7	1.21×10^7
RC06	7.20×10^7	4.14×10^7
RC07	8.38×10^7	3.24×10^6
RC08	3.20×10^7	2.98×10^7
RC09	1.88×10^8	4.80×10^2
RC10	8.38×10^7	4.64×10^6
RC13	N/A	N/A
RC14	0	1.98×10^7
RC15	0	3.42×10^4
RC18	2.76×10^7	7.02×10^6
RC20	0	4.4×10^6
RC22	4.37×10^7	9.4×10^6
RC25	1.89×10^7	7.1×10^5
RC26	0	5.3×10^3
RC28	1.48×10^6	1.4×10^5
RC30	5.39×10^7	1.6×10^3
RC31	2.65×10^7	3.0×10^6
RC33	1.11×10^7	2.7×10^6
RC34	4.67×10^6	2.2×10^5
RC39	0	9.8×10^5
RC40	2.69×10^7	1.2×10^5
RC42	0	4.0×10^6
RC44	3.23×10^7	9.5×10^5
RC45	1.19×10^8	0
RC51	6.36×10^7	1.9×10^7
RC54	1.87×10^6	1.7×10^6
RC55	0	3.7×10^5
RC56	3.31×10^7	5.9×10^6
RC60	4.64×10^7	1.3×10^7
RC61	5.95×10^7	2.0×10^7

Viremia (viral load) was assessed using a strain-specific qRT-PCR assay that amplifies highly conserved regions of ORF 7 from the krc1 and krc2 genomes (Bailey et al. 2014).

measures were greater for krc1 than for krc2 in every case, with a mean difference of 4.31×10^7 ($\pm 9.46 \times 10^6$ S.E.M) virions per mL ($P < 0.001$, paired T-test; Table 4.5).

Because we expected the strength of selection to be proportional to N_e, we predicted that strong signals of purifying selection would correspond to high viremia levels, increasing the magnitude of the difference between π_N and π_S. Indeed, for co-infected monkeys, median $|\pi_N - \pi_S|$ was 0.0134 for krc1 but only 0.00582 for krc2, a significant difference ($P = 0.0020$; Wilcoxon Signed Rank test). Further, according to the neutral theory, the amount of neutral polymorphism maintained in a population should be correlated with N_e. This prediction was supported by a positive correlation between $\pi_{S\text{-}NOL}$ and viremia overall ($r_S = 0.447$; $P = 0.0016$) and in krc1 ($r_S = 0.495$; $P = 0.0225$; Spearman's rank correlation); the correlation in krc2 alone was not significant.

4.2.5 Longitudinal Diversity Change and Mutation Rate Estimation

Unlike viruses such as HIV, which lead to the destruction of the immune system and eventual death of the individual host (Williamson et al. 2005), krc1 and krc2 maintain a persistent infection that is apparently asymptomatic. One co-infected monkey was sampled twice in this study, first as isolate RC05 on 11 February 2010, and next as isolate RC56 on 20 June 2012. Thus, the time elapsed between samplings was 860 days, or 2.36 years (2 years, 4 months, and 9 days). Such longitudinal data allow the estimation of viral mutation rates, making the reasonable assumption that non-overlapping synonymous sites evolve neutrally.

Mean between-population synonymous divergence ($\overline{d_S}$), the mean number of pairwise synonymous differences per synonymous site, can be used to estimate the

synonymous substitution rate r_S between two populations descended from a common ancestor over T generations, such that $r_S = \overline{d_S}/(2T)$ (Nei 1987). Under normal circumstances, if it can be assumed that synonymous mutations are not subject to selection, r_S is also equal to the per-site mutation rate v, since v is equivalent to the neutral substitution rate (Kimura 1983; Nei 1987). However, in the case of closely related populations sampled at two points in time, within-population variation as measured by π_S will contribute substantially to $\overline{d_S}$ and must be subtracted from the latter (Nei and Li 1979). In our case, one population is sampled at two time points separated by T years, such that the mutation rate can be estimated as:

$$\hat{v} = r_S = \frac{\overline{d_S} - \overline{\pi_S}}{T} \qquad \textbf{equation 4.1}$$

where $\overline{\pi_S} = (\pi_{S1} + \pi_{S2})/2$, and π_{S1} and π_{S2} are estimates of within-population synonymous nucleotide diversity at time points 1 and 2, respectively.

Because our analyses show that synonymous changes are constrained by overlapping ORFs, we used only non-overlapping codons in these analyses. We estimated $\overline{d_S}$ using viral variant data and custom Perl scripts based on SNPGenie subroutines, and used $\pi_{S\text{-}NOL}$ values depicted in Figure 4.2. This yielded mutation rate estimates of 8.02 X 10^{-3} and 6.88 X 10^{-3} per site per year for krc1 and krc2, respectively (Table 4.6). Thus, the mutation rate for krc1 is estimated to be about 1.16 times greater than that of krc2. Although the estimated synonymous substitution rates of RNA viruses

Table 4.6. Mutation rates for all ORFs of krc1 and krc2 as estimated from non-overlapping synonymous polymorphism.

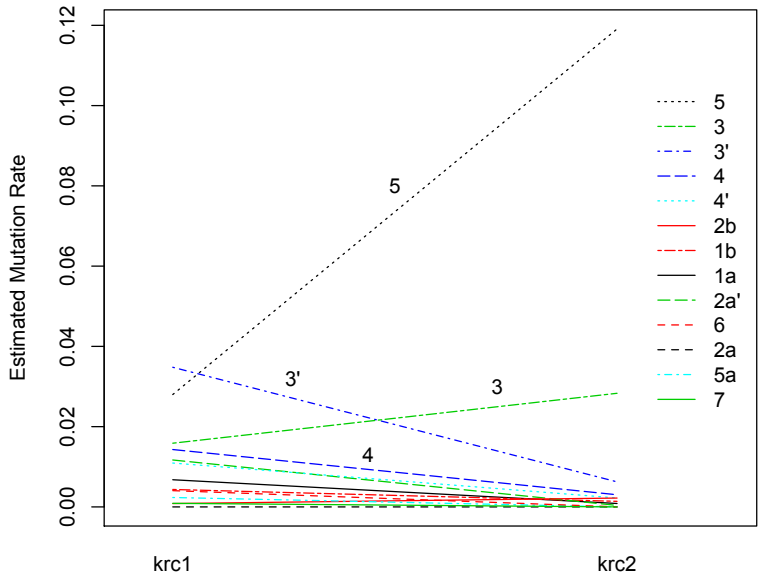
	krc1				krc2			
ORF	**Syn. Sites**	$\overline{d_S}$	$\overline{\pi_S}$	$\hat{v}$	**Syn. Sites**	$\overline{d_S}$	$\overline{\pi_S}$	$\hat{v}$
1a	1,402.98	0.0424	0.0265	6.77×10^{-3}	1,362.78	0.0193	0.0173	8.52×10^{-4}
1b	1,102.07	0.0396	0.0294	4.35×10^{-3}	1,068.00	0.0220	0.0187	1.39×10^{-3}
2a'	112.33	0.0516	0.0240	1.17×10^{-2}	91.13	0.0137	0.0128	3.74×10^{-4}
3'	101.30	0.1147	0.0327	3.48×10^{-2}	91.23	0.0244	0.0097	6.24×10^{-3}
4'	111.91	0.0560	0.0303	1.09×10^{-2}	94.86	0.0216	0.0164	2.22×10^{-3}
2a	6.33	0	0.0000	0	6.33	0	0	0
2b	50.59	0.0167	0.0147	8.35×10^{-4}	69.64	0.0242	0.0190	2.21×10^{-3}
3	62.05	0.0694	0.0320	1.59×10^{-2}	86.38	0.1057	0.0390	2.83×10^{-2}
4	91.97	0.0615	0.0278	1.43×10^{-2}	106.16	0.0195	0.0123	3.03×10^{-3}
5a	11.67	0.0230	0.0174	2.35×10^{-3}	9.50	0	0	0
5	144.12	0.0951	0.0292	2.80×10^{-2}	139.83	0.2858	0.0052	1.19×10^{-1}
6	120.13	0.0155	0.0060	4.05×10^{-3}	118.67	0.0023	0.0023	9.41×10^{-6}
7	84.17	0.0050	0.0029	8.88×10^{-4}	84.00	0	0	0
ALL	3,401.61	0.0452	0.0263	8.02×10^{-3}	3,328.51	0.0325	0.0163	6.88×10^{-3}

Mean numbers of synonymous sites and differences between RC05 and RC56 (2.36 years apart) were calculated using custom scripts and SNPGenie (Nelson et al. 2015). ORFs in krc2 differ in their start and stop positions between the two isolates, requiring them to be extracted and analyzed separately. ORFs which differed in length were then translated, aligned with ClustalW in MEGA7, and this alignment was imposed on the nucleotide sequence before analysis. Mutation rates were estimates using equation 4.1: $\overline{d_S}$ refers to mean between-isolate synonymous distance; $\overline{\pi_S} = (\pi_{S1} + \pi_{S2})/2$ where π_{S1} and π_{S2} are estimates of within-population synonymous nucleotide diversity at time points 1 and 2, respectively; $\hat{v}$ is the estimated mutation rate.

vary by at least 5 orders of magnitude, our mutation rate estimate falls in the center of the range of other estimates for members of *Arteriviridae*, which range from 5.20×10^{-3} to 6.12×10^{-2} (Hanada et al. 2004).

Our estimates suggest substantial mutation rate heterogeneity within the viral genomes (Figure 4.3). To see whether the differences were significant, we used factorial analysis of variance (ANOVA) with aligned codon units to test for main effects of the

(A)



(B)

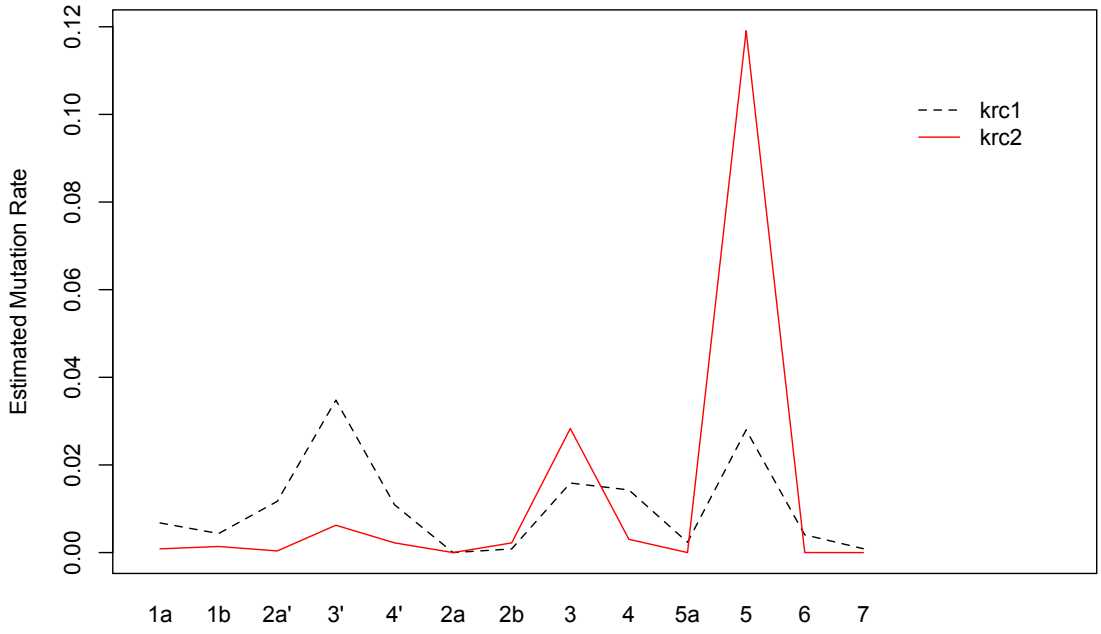


Figure 4.3. Interaction between virus and ORF mutation rate (per site per year) estimates. Although the overall mutation rate was estimated to be 1.16 times higher in krc2, this pattern saw a major reversal in ORF 5, and slight reversals in ORFs 3 and 2b. (A) ORFs with the top four mutation rates are annotated within the body of the chart. (B) ORFs are ordered 5' to 3' by start site from left to right.

virus (krc1, krc2), main effects of the ORF (1a, TF, 1b, 2a', 2b', 3', 4', 2a, 2b, 3, 4, 5a, 5, 6, and 7), and a virus-by-ORF interaction. The virus-by-ORF interaction was significant ($F_{12,8752} = 21.61$; $P < 0.0001$), suggesting that the difference in mutation rate between krc1 and krc2 is inconsistent across ORFs. ORFs 2b, 3, and 5 were the only ORFs that did not follow the overall pattern of krc1 > krc2. Of these, the greatest (and only significant) difference was observed in ORF 5, which had an estimated mutation rate 4.25 times higher in krc2 than krc1 ($P < 0.0001$; least squares means contrast with Tukey adjustment for multiple comparisons). The differences for ORFs 2b and 3 were not significant, being 2.64 and 1.78 times higher in krc2, respectively; however, it is noteworthy that these two ORFs neighbor one another in the genome. Of the remaining ORFs, in which the estimated mutation rate was higher in krc1 than in krc2, ORF 3' was the only significant difference observed, having a mutation rate 5.57 times higher in krc1 than krc2 ($P < 0.0001$; least squares Tukey adj.). The mutation rate of ORF 1a was also 7.94 times higher in krc1, but this difference was only marginally significant ($P = 0.0514$; least squares Tukey adj.).

Given that virus-by-ORF interaction was significant in the model, it follows that virus plays an important role in explaining mutation rate; however, after accounting for the interaction, the main effect of virus is not significant ($F_{1,8752} = 0.08$; $P = 0.7816$). On the other hand, the main effect of ORF remains highly significant ($F_{12,8752} = 46.42$; $P < 0.0001$). ORF 5 was significantly greater than all other ORFs ($P < 0.0001$ in all cases; least squares Tukey adj.). ORFs 3 and 3', both significantly less than ORF 5, did not differ significantly from one another, but did significantly exceed ORFs 2a', 1a, 1b, 6, 2b, and 6 ($P < 0.05$ in all cases; least squares Tukey adj.).

A simpler approach to estimate the viral mutation rate might be to calculate d_S between the consensus sequences of a viral population at two points in time. When this was done using our data, the mutation rate of krc1 was estimated to be 2.94 times greater than that of krc2 (1.35×10^{-2} and 4.58×10^{-3}, respectively; data not shown). This suggests that comparing consensus sequences alone is inadequate, as it can exaggerate population differences by failing to account for within-population diversity. Further, the mutation rate heterogeneity observed in the previous analysis was obscured when using the consensus approach, with ORF 3 having an estimated mutation rate only 1.36 times higher in krc2, and all other ORFs (including ORF 5) being highest in krc1. Importantly, subtracting within-host π_S from between-consensus d_S resulted in negative values for 4 of 13 (30.8%) ORFs in krc1, and for 9 of 13 (69.2%) ORFs in krc2, further evidencing the loss of data inherent in the taking of consensus sequences.

To consider a possible host immune mechanism for viral mutation rate heterogeneity, we analyzed all ORFs for enrichment in each of six preferred APOBEC3 motif targets (GG, TG, TGG, GGG, TGGG, GGGT; Ebrahimi et al. 2014). There was no significant correlation between the concentration of any of the motifs in an ORF and its estimated mutation rate (Spearman's rank correlation). In fact, while the concentration of motifs in the highly mutable ORF 5 was not particularly high, it was ORF 5a that had the highest concentration of all motifs except TG (second highest) and GGGT (absent), despite exhibiting a relatively low mutation rate (2.35×10^{-3} in krc1 and 0 in krc2).

The neutral theory predicts that most within-population diversity is selectively neutral. If this is the case, neutral polymorphism should accumulate over time as a consequence of mutation, which would be reflected by an increase in $\pi_{S\text{-}NOL}$. Over the

2.36 years elapsed between isolates RC05 and RC56, $\pi_{S\text{-}NOL}$ instead decreased from 0.0304 to 0.0223 in krc1, whereas it remained relatively constant in krc2, increasingly only slightly from 0.0161 to 0.0164. Neither change was significant (Wilcoxon Signed Rank tests). Thus, unlike other studies which demonstrate an increase in synonymous polymorphism over time (Nelson and Hughes 2015), our data fail to reject stasis. At the same time, the overall values of $\pi_{S\text{-}NOL}$ in these isolates were 0.0238 for krc1 and 0.0163 for krc2, a significant difference (P = 0.0135; Wilcoxon Signed Rank test, ORF unit). This is in keeping with neutral expectations, as the mean viremia of krc1 was 6.95 times greater than krc2 in these isolates, and $\pi_{S\text{-}NOL}$ is expected to correlate with N_e.

Li (1997) has noted that a strong correlation exists between nonsynonymous and synonymous substitution rates across many taxa, including RNA viruses, providing evidence that the mutation rate rather than positive selection drives evolutionary change. To test whether purifying selection is strong enough to eliminate the evolutionary signal of mutation pressure in krc1 and krc2, we measured the correlation between our estimated mutation rates and mean between-population d_N (time points 1 and 2) for non-overlapping regions of each ORF. All correlations were strong and significant: r_S = 0.830 overall ($P < 0.0001$), r_S = 0.767 for krc1 ($P < 0.0001$), and r_S = 0.909 for krc2 ($P < 0.0001$; Spearman's rank correlation). It is noteworthy that krc1, which has the larger N_e, has the weaker mutational signal. Additionally, we found that an ORF's mutation rate correlates significantly with its π_N at all coding sites (r_S = 0.579; P = 0.0020), as well as for overlapping (r_S = 0.501; P = 0.0091) and non-overlapping residues (r_S = 0.597; P = 0.0013; Spearman's rank correlation). This suggests mutation rate heterogeneity as a

possible mechanism for the π_N spatial pattern first observed by Bailey et al. (2014), in which 3'-proximal ORFs tend to have more nonsynonymous diversity.

Taken together, these observations indicate that, while mutation pressure drives evolutionary change in krc1 and krc2, the viruses are also not accumulating synonymous diversity at a rate rapid enough to be detected over the period of our study (2.36 years).

4.2.6 Gene Diversity and Tajima's D

Besides π, one measure that can be used to measure relative constraint is gene diversity (H), which can be compared at nonsynonymous, synonymous, ambiguous, and non-protein-coding sites. This measures the probability that two genomes randomly chosen from a population differ at the site of interest, and it can be estimated as $H = 1 - \sum_{i=1}^{n} x_i^2$, where x_i is the population frequency of the ith nucleotide variant and n is the number of variants observed (Li 1997). Most genes in most species display the greatest constraint at nonsynonymous sites, intermediate constraint in 5'- and 3'-UTRs, and the least constraint at synonymous sites, as determined by relative substitution rates (Li 1997; Graur and Li 2000; Hughes et al. 2003; Hughes, Packer, et al. 2005).

We used SNPGenie to estimate H at different SNP sites in krc1 and krc2. However, because we filtered out SNPs having an estimated frequency < 5% for quality control purposes, we expected that purifying selection would be reflected by a depletion in the relative number of nonsynonymous SNP sites but not in their mean frequency, since the majority of deleterious nonsynonymous SNPs would be expected to persist at frequencies << 5% given mutation-selection balance, and thus would not be detected by our methods (see Section 4.3).

Non-protein-coding regions of krc1 and krc2 are limited to a few hundred nucleotides at either end of the single-stranded RNA genomes. Coverage in these regions was often too low to allow SNP calling, but high-quality variable sites were present in krc2 isolates RC13 and RC14. These two viral populations contained 221 SNP sites that did not overlap multiple ORFs, with mean $H = 0.2476$ (± 0.0094 S.E.M.). Of these SNP sites, 100 (45.25%) were synonymous, 72 (32.58%) were nonsynonymous, 33 (14.93%) were non-protein-coding, and 16 (7.24%) were ambiguous. Approximately 75% of random mutations in coding regions are expected to be nonsynonymous (Nei 1975; Graur and Li 2000). Consistent with this, the average proportions of nonsynonymous and synonymous sites in these populations were 75.49% and 24.51%, respectively. Thus, the fact that only 41.86% of non-ambiguous coding SNPs are nonsynonymous is a significant deviation from the neutral expectation ($P < 0.0001$; Exact Binomial test), and is evidence that purifying selection has acted to reduce the number of such SNPs in these genomes. Results were similar when all populations of both viruses were included (data not shown).

Nonsynonymous SNP sites in krc2 RC13 and RC14 had the highest mean H of 0.3226 (± 0.0153 S.E.M.), followed by ambiguous sites with $H = 0.3050$ (± 0.0405), synonymous sites with $H = 0.2004$ (± 0.0117), and finally non-protein-coding sites with $H = 0.1994$ (± 0.0240). Different SNP site types exhibited significant differences in H ($P < 0.0001$; Kruskall-Wallis test), with H at synonymous sites being significantly less than that at nonsynonymous and ambiguous sites, but not significantly greater than non-protein-coding sites (Dunnett's multiple comparisons test; family error rate of 5%; random seed set to 61). Results were similar when all populations of both viruses were

considered, and differences between medians were even more pronounced than means (data not shown). Thus, non-protein-coding SNP sites exhibited the greatest evidence of purifying selection, while the nonsynonymous SNP sites exhibited the weakest. This implies that, for those nonsynonymous SNPs that *do* occur at frequencies > 5%, positive selection has acted to increase variant frequencies. Complementarily, purifying selection in these populations is sufficient to hold the frequencies of deleterious nonsynonymous mutations well below 5%.

Given that 58.14% of non-ambiguous coding SNP sites were synonymous when 24.51% were expected under neutrality, but that mean H was higher at the nonsynonymous SNP sites detected, we next asked how this would influence other population genetic estimators. One popular approach for detecting natural selection uses Tajima's D statistic (Tajima 1989). This measure compares two estimates of the population parameter θ, equivalent to $2N_e v$ for haploid populations, each of which is affected differently by selection. The first estimate is Watterson's θ_S, the number of SNP sites in a sample of sequences corrected for sample size, which ignores the frequencies of variants and is thus highly sensitive to rare alleles (Watterson 1975). The second estimate is Tajima's θ_π, equivalent to the average number of pairwise differences between sequences in a sample, which increases only slightly with the existence of rare alleles (Tajima 1983). Balancing selection and population admixture can increase the frequencies of rare alleles, leading to $\theta_S < \theta_\pi$ and a positive D (*i.e.*, few rare variants). On the other hand, purifying selection against deleterious variants or a population that is actively growing can result in a decrease in the frequency of deleterious alleles, leading to $\theta_S > \theta_\pi$ and a negative D (*i.e.*, many rare variants).

Given that our gene diversity estimates suggested purifying selection is extremely effective in reducing the frequencies of nonsynonymous deleterious alleles below our minimum allele frequency cutoff of 5%, we predicted that we would observe a positive *D* overall, falsely indicative of widespread overdominant selection. Indeed, all populations had positive *D* values ranging from 1.169 to 4.785, with a median of 2.798 (± 0.731 I.Q.R.) in krc1 and 2.049 (± 0.617) in krc2 (data not shown). These results demonstrate that a comparison of the number of nonsynonymous and synonymous segregating sites, and/or a nucleotide diversity analysis, is necessary for detecting purifying selection in quality-filtered pooled-sequencing results when overdominant selection is taking place in any regions of the source population's genome.

4.3 Discussion

Estimation of population genetic parameters using pooled-sequencing viral data allows unprecedented insight into their within-host evolution. In this study, we go beyond mere comparisons of π_N and π_S to explore the effects of selection in different regions of SHFV viruses krc1 and krc2. As is true for almost all organisms studied to date, π_N was significantly less than π_S at the genome level for both viruses. While it has sometimes been hypothesized that ORF 3' is silent (Godeny et al. 1998), our evidence suggests that this is not the case, since this ORF also exhibits strong evidence of purifying selection, which should only occur if it is expressed. Interestingly, ORFs 4' and 3 are expressed only in small amounts during virus replication, which might lead us to hypothesize that they are subject to the most relaxed purifying selection (ORFs 3 and 5 in Godeny et al.

1998). However, $\pi_N < \pi_S$ was significant for both ORFs in krc1 and for ORF 4' in krc2, implying important functional constraint.

All ORFs in these viruses have regions which overlap other ORFs, although the number of overlapping codons was dramatically lower in ORFs 3, 2b, and 4 of krc2. We predicted that purifying selection would be stronger at overlapping residues, leading to a reduction in π_N at these as compared to non-overlapping sites, *i.e.*, $\pi_{N\text{-}OL} < \pi_{N\text{-}NOL}$. Contrary to this prediction, $\pi_{N\text{-}OL}$ was significantly greater than $\pi_{N\text{-}NOL}$ in both viruses. Moreover, $\pi_{S\text{-}OL}$ was significantly less than $\pi_{S\text{-}NOL}$ in both viruses.

One possible explanation for the lack of constraint in overlapping regions is that they are less functionally important. However, some viruses are known to contain epitopes in overlapping regions (Hughes et al. 2001), suggesting the more likely possibility that $\pi_{N\text{-}OL}$ might be elevated in krc1 and krc2 because their overlapping regions contain epitopes undergoing ovedominant selection for immune escape. In particular, π_N significantly exceeded π_S in overlapping residues of krc1 ORFs 4' and 5, and krc2 ORFs 4 and 5. If true, the observation that $\pi_{S\text{-}OL}$ was significantly less than $\pi_{S\text{-}NOL}$ in both viruses could also be explained by the fact that many synonymous changes in overlapping regions would be nonsynonymous in the overlapping ORF, leading them to experience purifying selection that would decrease their frequencies. Finally, overdominant selection on overlapping residues could help explain why there was no significant correlation between an ORF's π_S and its proportion of overlap in either virus.

We identified candidate epitopes as nonsynonymous peaks, 9-codon sliding windows in which π_N exceeded 0, the window's π_S, and the ORF's π_S. These windows were then concatenated and end-trimmed, yielding 12 peaks across 6 ORFs of krc1 and 8

peaks across 5 ORFs of krc2 (Table 4.2). Importantly, peak 1 in ORF 5 of both viruses overlapped ORF 5a, explaining why the overlapping residues of 5a exhibit such high levels of nucleotide diversity. Additionally, ORFs 3 and 5 each contained two nonsynonymous peaks in both viruses, explaining the previous identification of these ORFs as ones likely to be under positive selection (Bailey et al. 2014). It is interesting to note that these are also the two ORFs in which the mutation rate is substantially higher in krc2.

Peaks in both viruses tended to occupy similar ORFs and positions within those ORFs. They were indeed located disproportionately in overlapping ORF regions, supporting our hypothesis that the elevated $\pi_{N\text{-}OL}$, diminished $\pi_{S\text{-}NOL}$, and lack of correlation between proportion of overlap and π_S in these viruses are due to the presence in overlapping regions of epitopes undergoing overdominant selection for immune escape. However, of the 4 ORFs exhibiting significant $\pi_{N\text{-}OL} > \pi_{S\text{-}OL}$, exclusion of the peak residues only caused the ratio to reverse in one. Thus, overdominant selection may only be a partial explanation. On the other hand, our nonsynonymous peak criteria may have been too stringent to identify every epitope, or positive selection too weak to produce a sufficiently strong signal.

Because overlapping regions disproportionately house nonsynonymous peaks, the constraint imposed by overlapping ORFs was not apparent from straightforward analyses of nucleotide diversity at the genome level or in the majority of the ORFs of either viruses. We therefore sought evidence for purifying selection in these regions using an alternative approach that controls for the effects of overdominant selection. Of all possible nonsynonymous mutations that occur in overlapping ORFs, approximately 64%

are expected to result in a nonsynonymous change in the alternative ORF, assuming the frames differ (always the case in SHFV). However, only approximately 30% of nonsynonymous changes observed in peak-containing ORFs fit this expectation, the majority being synonymous in the alternative ORF. Thus, purifying natural selection acting on an overlapping region of one ORF in these viruses can constrain the nonsynonymous changes it undergoes, such that they will disproportionately be synonymous in the alternative ORF. This pattern was highly significant in the majority of ORFs, both in the peak residues and the remainder of the overlap. The only exception to this pattern was the one instance in which the remainder of one peak coincided with a peak in the alternative frame. The most straightforward explanation is that nonsynonymous changes in the alternative ORF were being favored by overdominant selection. Moreover, all other remainder residues, which did not overlap peaks in alterative ORFs, were even more constrained than the peak residues in terms of the accepted nonsynonymous changes. Thus, the existence of a peak does act to promote nonsynonymous polymorphisms in alternative ORFs, but purifying selection is still able to significantly constrict which changes these will be.

Having established that purifying selection constrains nonsynonymous mutations to disproportionately cause synonymous mutations in overlapping ORFs, an interesting implication immediately presents itself. If selection greatly limits what nonsynonymous mutations are able to be used by the virus to escape immune recognition, this means that the options for nonsynonymous escape mutations favored by selection in overlapping regions are to some extent predictable. This is especially true given the high mutation rates of these viruses, which we estimated as 8.02×10^{-3} and 6.88×10^{-3} per site per year

for krc1 and krc2, respectively. Given their high viremia, it is almost certain that mutations produce all possible SNPs that are 1 nucleotide removed from the viral consensus sequence of a within-host population each viral generation. All of these mutations are able to be tested by natural selection (with the vast majority of nonsynonymous ones being deleterious, as evidenced by $\pi_N < \pi_S$). Thus, for a given viral genome sequence, knowledge of which nonsynonymous changes lead to synonymous changes in overlapping ORF may improve our understanding of epitope evolution and its likely trajectories. Furthermore, one reason that peaks disproportionately map to overlapping regions might be that host immune systems have more success targeting these regions on account of their being more constrained. In other words, more epitopes may exist in overlapping regions simply because they have long ago been lost in non-overlapping regions.

Besides nonsynonymous polymorphism, our data strongly suggest the jettisoning of genomic material as one mechanism by which these viruses can achieve immune escape. ORF 3' contained 2 nonsynonymous peaks in krc1 that were absent in krc2, and also contained two 9-codons deletions in krc2 as compared to krc1. While the coordinates of these deletions did not match those of the peaks in krc1, they may have contributed to immune escape by altering the configuration of the epitope within the gene product. Alternatively, the epitopes may have changed positions in the ORFs before being jettisoned. For krc1 ORF 4' peak 2, the first 9 codons (the majority of the peak) were deleted in krc2. Between-virus alignment gaps were also prevalent in krc2 ORF 5 peaks 1 and 2, both peaks of which are also present (but without sequence gaps) in krc1.

Other evidence for the role of indels in immune escape comes from within-population variation. All peaks containing indels exhibited alignment gaps in the majority of isolates, and these often differed in location, suggesting separate underlying mutations. For example, in krc1 ORF5, all isolates exhibiting peak 2 contained both alignment gaps and nonsynonymous polymorphism. Interestingly, the one isolate which does not contain nonsynonymous polymorphism within this peak is also the only isolate which has no alignment gaps. It is noteworthy that the only ORFs differing in size by more than 2 central codons among populations of the same virus—ORF 5 in krc1 and ORFs 3 and 5 in krc2—all contained nonsynonymous peaks. ORFs 3, 4, 5a, and 7 of krc1 and ORFs 2b, 6, and 7 of krc2 also differed in length among isolates, but these differences were observed either entirely at the end of the ORF or else involved no more than 2 codons.

Besides suggesting indels as a major mechanism by which immune escape can occur, these results complementarily suggest that most indels which become common in these viruses do so as a result of overdominant selection. This is in keeping with the prediction that many beneficial mutations will be loss-of-function, since it is easier for mutation to damage or deactivate genomic material than create it (Hughes 2007; Hughes 2012). It is also noteworthy that only one peak contained indels in an overlapping region (krc1 ORF 3 peak 2), involving a median of only 1 codon. Thus, overlapping residues may be less likely to contain indels, presumably on account of their heightened functional constraint.

According to the neutral theory, natural selection is expected to act with greater efficacy in larger populations. Our data support this prediction, as $|\pi_N - \pi_S|$ was greater in krc1, which had higher viremia, than in krc2. We also support the neutral prediction that

larger populations maintain more neutral diversity, as measured by $\pi_{\text{S-NOL}}$, although this did not hold when krc2 isolates were considered alone. Neutral diversity can be used to estimate the population parameter θ, which is proportional to N_e and the mutation rate, v. We would expect in the haploid case that $\pi_{\text{S-NOL}} = 2N_e v$ (Nei and Kumar 2000; Lynch 2007a). When comparing two viruses such as krc1 and krc2, we would further expect that:

$$\frac{\pi_{\text{S-NOL1}}}{\pi_{\text{S-NOL2}}} = \frac{2N_{e1}{}^{v_1}/_{t_{G1}}}{2N_{e2}{}^{v_2}/_{t_{G2}}} \quad \textbf{[equation 4.2]}$$

where t_G is generation time. Thus, a significant difference in $\pi_{\text{S-NOL}}$ between two populations may be due to differences in N_e, v, or t_G.

In our longitudinal isolates, viremia is 6.6 times higher in krc1 on average; thus a larger N_e helps to explain the higher $\pi_{\text{S-NOL}}$ observed in krc1. Further, we also estimate that v is 1.2 times higher in krc1. However, $\pi_{\text{S-NOL}}$ itself was only 4.8 times higher in krc1. To address this higher-than-expected N_e ratio, we might first make the reasonable assumption (supported by our statistical analyses) that v is the same in both viruses. To this end, evidence suggests that differences in RNA virus mutation rates per unit time are due primarily to differences in replication rates (*i.e.*, t_G) rather than to differences in replication error (*i.e.*, v) (Hanada et al. 2004). We might then conclude that the virus with the higher $\pi_{\text{S-NOL}}$ (krc1) has a higher N_e or a shorter t_G. In fact, N_e and t_G may not necessarily be independent, as faster (shorter) replication times may contribute directly to the higher viremia observed in krc1. However, the ratio implied by viremia alone suggests that $\pi_{\text{S-NOL}}$ should be even greater in krc1 than we observe. One explanation for

this might be that the relationship between viremia and N_e is negatively allometric, *i.e.*, it levels off. On the other hand, t_G might actually be larger in krc1, acting to decrease the $\pi_{S\text{-}NOL}$ ratio. Whether the values of $\pi_{S\text{-}NOL}$ are due entirely to N_e, or if v and t_G also play a role, are questions that should be addressed in future longitudinal studies.

Higher v can increase the value of π_N by exerting mutation pressure that can overwhelm purifying selection. This held true in our study, as π_N correlated significantly with v in both viruses. As expected from N_e, this correlation was weaker in krc1 (r_S = 0.767) than krc2 (r_S = 0.909). Interestingly, a factorial ANOVA using codons as independent mutational units supported the hypothesis that krc1 and krc2 experience within-genome mutation rate heterogeneity. Because 3'-proximal ORFs tended to have higher mutation rates, such heterogeneous mutation pressure could help explain why these ORFs also have higher π_N. In particular, ORFs 3 and 5 both had the two highest estimated mutation rates and the two highest π_N values in both viruses. Most indels also occur in 3'-proximal ORFs, which we never observed to occur before ORF 2b in either virus, being limited to the final ~20% of the genome's 3' end. Thus, besides having a higher mutation rate, the 3' regions of these viruses are enriched in nonsynonymous changes, nonsynonymous peaks, and indels.

Unfortunately, analyses for enrichment in known APOBEC3 motif targets were not significant, and a likely mechanism for mutation rate heterogeneity remains elusive. Other studies examining such heterogeneity between loci have identified the phenomenon at much higher levels of resolution, *e.g.* 200 kb, which greatly exceed SHFV genome length (Ness et al. 2015). Although krc1 and krc2 are non-segmented viruses in which various subgenomic RNAs containing different ORFs are produced for gene expression

purposes, only full minus-strand RNA transcripts including all ORFs are used for replication, implying no biological differences between ORFs in the replication process. On the other hand, the genome's involvement in RNA secondary structures might contribute to mutability (Holmes 2009). Finally, the fact that we estimated a mutation rate of 0 for 4 of the 26 virus/ORF combinations tested may suggest that 860 days is insufficient sampling time to properly distinguish differences between ORFs in every instance.

Future longitudinal studies with SHFVs should seek to model potential interactions between virus and ORF to confirm or deny our mutation rate findings. If correct, the competitive exclusion principle could be used to direct investigation of whether krc1 and krc2 occupy distinct niches within their red colobus hosts (as suggested by the fact that viremia and co-infection are independent), and whether predation by the immune system keeps viremia well below the carrying capacity of the within-host micro-environment (Hardin 1960; den Boer 1986; Nowak and May 2000). These questions are more than theoretical, as defining such carrying capacities could help to decipher the evolutionary dynamics of emergent viruses during host-switching (*e.g.*, for influenza; Moncla et al. 2016).

It is critical to note that our estimate of v in no way depends on π_N. Nor is the correlation between these two just a restatement of the correlation between π_S and π_N, since within-population π_S is explicitly taken into account in our v estimator (equation 4.1). Epigenetic mechanisms would also not greatly alter our conclusions, because in either case the result is selective immune escape. Some might point out that it is possible for synonymous mutations to influence protein structure, making them subject to natural

selection (Komar 2007; Hunt et al. 2009; Kimchi-Sarfaty et al. 2016). Indeed, there are examples of synonymous mutations which are subject to strong selection, *e.g.*, in the ribosomal S20 gene of *Salmonella enterica* (Knöppel et al., *in press*). To the extent that this occurs, nucleotide diversity analyses will constitute a conservative test for purifying selection, a signal that is nevertheless ubiquitous. Yet $\pi_N < \pi_S$ for both krc1 and krc2, indicating that synonymous sites are relatively unconstrained as compared to nonsynonymous sites in these viruses. Moreover, purifying selection against synonymous changes would deflate our estimate of the mutation rate, which nevertheless falls in the center of previous estimates for Arteriviruses. It is also difficult to imagine overdominant selection favoring a great many synonymous changes, since it is unable to act even on nonsynonymous changes with perfect efficacy.

In addition to π, estimates of gene diversity (H) and Tajima's D can be used to detect the effects of natural selection. Like π, these parameters suggest that synonymous sites are relatively unconstrained as compared to nonsynonymous sites, the former of which are comparable to non-protein-coding sites in most systems (Li 1997; Graur and Li 2000). However, it is questionable whether they can be informative when used with pooled-sequencing data. Of the non-ambiguous SNP sites observed, only 41.86% were nonsynonymous as compared to the neutral expectation of 75.49%, strongly supporting the ubiquity of purifying selection in the viral genomes. However, contrary to expectation, H was significantly higher at nonsynonymous than at synonymous SNP sites, a signature normally indicative of widespread positive selection that is in conflict with our other analyses. Moreover, Tajima's D, which relies on the number of SNP sites, also indicated widespread positive selection.

This apparent contradiction can be easily explained by mutation-selection balance and the quality control measures currently necessary for pooled-sequencing SNP calling. Site-directed mutagenesis experiments with RNA viruses have allowed unprecedented insight into the distribution of their mutational fitness effects, with fitness generally measured as a change in replication rate as compared to an ancestral viral genome within a laboratory cell medium. The fraction of lethal mutations ranges from 28.6% to 40.9%, while those mutations that are not lethal have an average (deleterious) fitness effect of -0.103 to -0.132 (Sanjuán 2010). Thus, the distribution of mutational fitness effects for RNA viruses appears bimodal, with most mutations being either lethal or slightly deleterious. If it could be shown that purifying selection against most deleterious mutations is such that they are expected to segregate at frequencies far below our SNP calling cutoff of 5%, this would explain our finding that mean gene diversity is higher for nonsynonymous than for synonymous SNP sites in our variant data.

Population genetics theory can be used to derive the equilibrium frequency of deleterious variants at mutation-selection balance. However, this requires a mutation rate per generation rather than per year. One approach for determining viral generation time might compare estimates of mutation rates per site per replication (viral generation) to synonymous substitution rates per site per year. Using data from the Arterivirus causing porcine reproductive and respiratory syndrome, this yields an estimate of 5.23 hr (*i.e.*, 1,675.7 generations per year) (Hanada et al. 2004). An alternative approach examines the virus' time to plateau in one-step growth curves (Rafael Sanjuán, personal communication). Caì et al. (2015) have recently shown that SHFV titers peak in MA-104 kidney cells at 36-60 hr, with titers beginning to fall by 72 hr. Estimates for peak time in

other RNA viruses range from 10-48 hr (Llewellyn et al. 2002; Mishra et al. 2010; Pliaka et al. 2011). However, generation time depends not only on intracellular viral replication, but also upon cellular exit and infection of new cells. Given that budding is thought to take 24-48 hours post-infection for Arteriviruses and other RNA viruses (Stueckemann et al. 1982; Bächi 1988), we take 96 hr as a conservative estimate of generation time (*i.e.*, 0.25 generations per day). This yields mutation rates per site per generation of 8.79×10^{-5} and 7.54×10^{-5} for krc1 and krc2, respectively. We note that extended latency during persistent infection would reduce the number of generations, thereby causing our estimate of the mutation rate per generation to be an underestimate. However, the persistently high viremia observed in all examined SHFV-positive RC monkeys makes this unlikely.

Given a mutation rate per generation, the equilibrium frequency of deleterious alleles can be calculated for haploids or fully dominant alleles as $q = u/|s|$, and is almost independent of population size (Crow and Kimura 1970). Given that krc1 and krc2 participate in persistent, asymptomatic infection, it is reasonable to assume that approximate equilibrium has been reached. Thus, given the estimated mutation rates per site per generation, and assuming a mean deleterious fitness effect in the range -0.103 to -0.132, we can estimate that the expected value of q should be bounded at the lower end by 0.057% (krc2) and at the higher end by 0.085% (krc1). Thus, we would indeed expect the great majority of deleterious mutations to persist at frequencies far below our 5% quality control cutoff, explaining the paucity of low-frequency nonsynonymous mutations in our dataset, and the gene diversity and Tajima's D estimates that follow. This also suggests that non-protein-coding SNP sites, which had the lowest observed H, are indeed under purifying selection, but that this is not so strong as the selection acting

against nonsynonymous SNPs, since their frequency at mutation-selection balance sits well above 5%.

Note that, if mutation/selection balance were to center on a mean frequency of 5%, the mutation rate would need to be 5.15 X 10^{-3} per site per generation, which is two orders of magnitude greater than our and others' estimates. Should episodes of latency occur, the mutation rate would need to be even higher to meet this condition. Thus, we feel confident in our interpretation that the great majority of nonsynonymous deleterious mutations likely segregate at frequencies far below 5% in these viral populations.

These results suggest an important caveat for population genetics estimates based on pooled-sequencing data (Futschik and Schlötterer 2010), namely, that estimates of parameters such as *H* and Tajima's *D*, which rely on knowledge of low-frequency segregating sites, may not have a straightforward interpretation. In our case, a traditional interpretation of Tajima's *D* would erroneously indicate the prevalence of overdominant selection but not purifying selection. Paradoxically, this result actually arises from the extreme efficacy of purifying selection in viral populations, which keeps the frequencies of deleterious variants low. Thus, it must be recognized that whole genes and populations are not simply "under purifying selection" or "under positive selection"; rather, genes and genomes are subject to a complex interplay of various evolutionary forces, the signals of which may be obscured depending on the level of genomic resolution under study. In our case, the proportion of SNP sites which were nonsynonymous reflected widespread purifying selection, while the high value of *H* at those sites resulted from a relatively small number of nonsynonymous peaks in the genome. Until quality control measures for pooled-sequencing can be developed which allow us to detect the majority of rare

variants with confidence, such population parameter estimates will require careful interpretation.

4.4 Materials and Methods

Blood samples were collected from all animals in previous studies following the guidelines of the Weatherall Report on research using non-human primates, as described in Bailey et al. (2014). Briefly, red colobus (RC) monkeys were sampled between 2/5/2010 and 7/22/2012 in Kibale National Park, Uganda (centroid 0.50°N, 30.40°E). Thirty (30) RC isolates were SHFV-positive of the 60 examined (50%); 23 were infected with krc1, 27 with krc2, and 21 were co-infected. Blood samples were obtained following the use of anesthesia, after which animals were returned to their social group. Blood was separated using centrifugation, frozen, and returned to the USA Wisconsin National Primate Research Center for study. For each animal, 1mL of blood plasma was filtered, viral RNA isolated, and DNase treatment performed. Quantitative RT-PCR was used to estimate viremia (viral RNA copies per mL of blood plasma) using highly conserved regions of ORF7.

As described in Bailey et al. (2014), pooled cDNA was synthesized using random hexamers and deep sequenced using Illumina MiSeq (Illumina, San Diego, CA, USA). Low-quality (< Q25) and short (< 100 bp) reads were filtered and *de novo* assembly performed using a customized method to minimize cross-mapping of krc1 and krc2 reads in co-infected animals, yielding < 0.2% cross-mapping, in CLC Genomics Workbench 5.5 (CLC bio, Aarhus, Denmark). The resultant population consensus sequences correspond to GenBank accession numbers KC787607-KC787658. Coverage (reads per

site) ranged from 119 to 19,115 (mean 5,654) for krc1 and from 94 to 6,613 (mean 2,264) for krc2. Geneious R5 (Biomatters, Auckland, New Zealand) was used to call single-nucleotide polymorphisms (SNPs) with a minimum coverage of 100 and a minimum allele frequency of 5%.

All custom scripts were written in Perl or R, figures were made in R and modified in PowerPoint, and statistical analyses were performed in R version 3.0.2 (R Core Team 2013; http://www.R-project.org/). Measures of spread were reported as S.E.M. (standard error of the mean) or I.Q.R. (interquartile range) as appropriate. When relevant, tests were two-sided. Exact Binomial tests used `stats:binom.test()`; Fisher's Exact tests used `stats:fisher.test()`; Kruskall-Wallis tests used `stats:kruskal.test()`; Dunnett's test used `multcomp:glht()` with `linfct = mcp(factor.values = "Dunnett")`; two-sample T-tests used `stats:t.test()` `paired = F`, while paired T-tests used the same function with `paired = T`; Wilcoxon Sign tests used `stats:wilcox.test()` with `paired = F`, while Wilcoxon Signed Rank tests used the same function with `paired = T`; and correlation tests used `stats:cor.test()` with `method = "spearman"` for Spearman's rank correlation. When outcomes depended on random number seeds, the seed was chosen using `base:sample(1:1000,1)` and set with `base:set.seed()`.

Nucleotide diversity at nonsynonymous and synonymous sites (π_N and π_S, respectively) was calculated using a new method for pooled NGS data (Nelson and Hughes 2015) based on that of Nei and Gojobori (1986) using SNPGenie version 1.2.2 (Nelson and Hughes 2015; Nelson et al. 2015; https://github.com/hugheslab/snpgenie).

This approach provides an accurate estimate when the number of substitutions per site is ≤ 0.1 (Nei and Kumar 2000). When comparing viral populations between or within hosts, ORF sequences were extracted from the genome sequence using a custom script, translated, and aligned at the amino acid level using the CLUSTAL algorithm in MEGA7 (default settings; Kumar et al. *in press*; Tamura et al. 2013). This alignment was then imposed on the nucleotide sequence.

In order to estimate viral (meta-population) π_N and π_S in sliding windows of 9 codons, we concatenated and aligned results from all isolates. Mean coverage estimates were used for multi-nucleotide variants. Nonsynonymous peaks, regions likely to be under overdominant positive selection, were identified conservatively as windows in which π_N exceeded 0, the window's π_S, and the ORF's π_S.

In order to estimate mean between-virus d_N and d_S for longitudinal isolates, a representative sample of genome sequences was generated with size equal to the viral population's maximum NGS coverage depth for a single polymorphic site: n = 3,244 for krc1 host RC05; n = 2,914 for krc1 host RC56; n = 962 for krc2 host RC05; and n = 962 for krc2 host RC56. Sequences were generated using custom Perl scripts based SNPGenie, which randomly distributed the observed variants throughout the sequence sample with frequencies equal to those observed in the SNP calling reports. MEGA7 software was unable to handle these sample sizes (Kumar et al. *in press*; Tamura et al. 2013). The mutation rate was then estimated using equation 4.1.

To estimate d_S between consensus sequences, numbers of nonsynonymous and synonymous sites were calculated using the Nei-Gojobori method in MEGA7 (Kumar et al. *in press*; Tamura et al. 2013). Since krc2 isolates differed in their non-protein-coding

leading and trailing material, these regions were manually removed. SNPGenie codon results for all non-overlapping sites were extracted using Perl scripts.

We modeled the estimated mutation rate using a factorial analysis of variance (ANOVA) model with virus, ORF, and virus-by-ORF interaction terms. This was accomplished by building a linear model with the R `stats:lm()` function with the `contrasts` option set to `contr.sum` for both factors, and then using the `car:Anova()` function with `type = 3` to perform type III (drop-one) F tests for each term. Results were verified in SAS. Because different virus/ORF combinations contained differing numbers of codons, the ANOVA was unbalanced, necessitating the use of least squares means to perform multiple comparisons. This was accomplished using the `lsmeans:lsmeans()` function with the `tukey` argument for all model terms. Results were verified in SAS.

Gene diversity (*H*) was calculated using the methods of Hughes et al. (2003). Ambiguous SNP sites were defined as those having both nonsynonymous and synonymous variants as compared to other viruses in the same isolate. Tajima's *D* was calculated for each viral population as:

$$D = \frac{\Pi - (S/a_1)}{\sqrt{\mathrm{V}(\Pi - (S/a_1))}} \qquad \textbf{[equation 4.3]}$$

where Π is the average number of pairwise differences between sequences in the isolate, S is the number of segregating sites, a_1 is a correction factor for sample size, and the denominator is the standard error of the difference. The latter two were computed as

described by Tajima (1989), with minimum coverage used as the number of sequences being surveyed (sample size), using 1,000 as the upper limit.

4.5 Acknowledgements and Funding

The authors thank the University of Wisconsin Department of Pathology and Laboratory Medicine and the Wisconsin National Primate Research Center (WNPRC) for the use of its facilities and services. C.W.N. thanks Meredith Yeager (NCI) for helpful discussion and feedback, and James Hussey (UofSC) for statistical training and advice. This work was funded by the joint NIH-NSF Ecology of Infectious Diseases program and the UK Economic and Social Research Council (TW009237) and the Wisconsin Partnership Program through the Wisconsin Center for Infectious Diseases. This study was made possible in part by grants from the National Institutes of Health National Center for Research Resources (P51RR000167) to the Wisconsin National Primate Research Center (WNPRC), University of Wisconsin-Madison. Biological research was conducted in part at a facility constructed with support from Research Facilities Improvement Program grant numbers RR15459-01 and RR020141-01. A.L.B. performed this work with support from the University of Wisconsin's Medical Scientist Training Program (MSTP) (grant T32 GM008692) and a National Research Service Award (NRSA) through the Microbes in Health and Disease (MHD) training program at the University of Wisconsin (T32 AI55397). C.W.N. performed this work with support from NSF Graduate Research Fellowship DGE-0929297, the University of South Carolina (USC) Presidential Fellowship, and the USC Department of Biological Sciences Kathryn Hinnant-Johnson,

M.D. Memorial Fellowship. The funders of this research had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

4.6 Author Contributions

A.L.H. and C.W.N. conceived the study questions. M. Lauck and S.D. Sibley performed sequencing experiments. M.L. and A.L. Bailey produced *de novo* sequences and performed variant calling. C.W.N. performed data analyses and wrote the manuscript.

CHAPTER 5

CONCLUSION

Among students of evolutionary biology, there has been a strong tendency to claim that [some] population genetical parameters will never be known accurately and therefore theories which contain such parameters are of little use. I take the opposite view; these parameters have to be investigated and measured if we really want to understand the mechanism of evolution at the molecular level. Can astronomers and cosmologists claim that theories which contain various astronomical parameters should be avoided because such parameters are difficult to estimate accurately? This reminds me of an aphorism, which I understand is due to Galileo, and which in effect says: what we can measure we should measure; what we cannot measure at present, we should endeavor to make measurable… what is important in science is to find out the truth.

— Motoo Kimura (1983)

Having defined the population genetic factors determining the selective potential of biological populations, we saw in Chapter 1 that the high reproduction rates and population sizes of RNA viruses make them especially amenable to the study of positive natural selection. Specifically, simian hemorrhagic fever viruses have large population sizes, sometimes on the order of 10^8 per mL in their monkey hosts (Chapter 4), and likely exhibit enormous replication rates, with other RNA viruses having burst sizes on the order of 10^4 per cell (Chen et al. 2007). The appearance of adaptive mutations and burden of deleterious mutations are thus the rate-limiting phenomena in adaptive RNA virus evolution.

Next, we saw in Chapter 2 that next-generation sequencing (NGS) data using representative pooled samples—the genomes of multiple biological entities combined in one sequencing run—can be used to call single nucleotide polymorphisms (SNPs), the frequencies of which are reliable estimates of allele frequencies in the source population (Futschik and Schlötterer 2010). Specifically, SNP frequencies allow the use of simple methods for estimating population genetic parameters, including nucleotide diversity (π) and gene diversity (H) at nonsynonymous and synonymous sites. Determining the values

of these parameters at different scales, from the single nucleotide to the whole genome, allows numerous evolutionary hypotheses to be tested, including the prevalence of positive (Darwinian) and/or negative (purifying) selection and the applicability of the neutral theory (Kimura 1983) to within-host virus dynamics.

In Chapter 3 we introduced SNPGenie, a new bioinformatics tool, written in Perl, that can be used to automate the estimation of the aforementioned parameters. In the time since its inception, SNPGenie has been used to study the transmission of H1N1 (Moncla et al. 2016) and H5N1 (Wilker et al. 2013) influenza in ferrets, conditional immune escape in simian immunodeficiency virus (Gellerup et al. 2016), evolution of Nod-Like Receptor resistance genes of the wild tomato *Solanum pennelli* (Stam et al., *in press*), and natural isolates of Arteriviruses in red colobus monkeys (Bailey et al. 2014; Nelson and Hughes 2015; Chapter 4) and Arteriviruses, pegiviruses, and lentiviruses in African Green Monkeys (Bailey et al., *in press*). It has also been improved to accept the standard SNP data format, the variant call format (VCF), and can analyze both '+' and '-' strands for double-stranded genomes, which may be of use in the study of overlapping bidirectional genes.

Chapter 4 takes advantage of the most recent advances implemented in SNPGenie to address questions about red colobus (*Procolobus rufomitratus tephrosceles*) Arterviruses simian hemorrhagic fever virus (SHFV)-krc1 and SHFV-krc2 which were previously prohibitive. We first show through comparisons of nonsynonymous and synonymous π (π_N and π_S, respectively) that the genomes of both viruses experience widespread purifying selection, confirming previous results for SHFV (Bailey et al. 2014) and most other viruses (Holmes 2009; Nelson and Hughes 2015). Regarding the

constraint imposed by overlapping open reading frames (ORFs), we find that overlapping regions indeed experience constraint in terms of what nonsynonymous variants are acceptable, namely, disproportionately ones which cause *synonymous* changes in the alternative ORF. However, this signal is not detectable on the genome scale, because we find that the majority of nonsynonymous peaks—sliding windows in which π_N exceeds π_S within both the ORF and the window—map disproportionately to overlapping regions, evidencing overdominant selection (heterozygote advantage) (Hughes and Nei 1988).

We further show that, when populations are sampled as natural isolates from the same host at distinct time points, straightforward population genetic theory can be adapted to NGS data to estimate mutation rates. When this was done, we obtained estimates of 8.02×10^{-3} mutations per site per year for SHFV-krc1 and 6.88×10^{-3} mutations per site per year for SHFV-krc2, falling in the center of previous estimates for Arteriviruses (Hanada et al. 2004). Statistical analyses suggest the possibility of mutation rate heterogeneity in the SHFV genome, with 3'-proximal ORFs exhibiting higher rates. If true, this could help to explain the high π_N of these ORFs, as well as their enrichment in nonsynonymous peaks and insertions/deletions. Unfortunately, a mechanism for this heterogeneity remains elusive.

Population parameters such as gene diversity and Tajima's *D* are alternatives to π for detecting the effects of selection. However, when applied to our NGS data for SHFV, both yield the conflicting result that overdominant positive selection rather than purifying selection is most widespread. We show that, given the plausible range of within-host replication rates and our estimated mutation rates, mutation-selection balance would be expected to maintain the equilibrium frequency of a typical nonsynonymous deleterious

allele well below 5%. This is below our minimum frequency quality control cutoff, explaining the contradictory implications. As a result, we suggest that parameters which rely on the detection of rare variants may not be of much use until pooled NGS methods are improved. At present, π—which relies neither on linkage nor rare variants—is the best choice.

A few obvious avenues for continuing this research present themselves. First, it is critical that larger samples of longitudinal data be analyzed, so that more sophisticated statistical models can be brought to bare on mutation rate estimations. This will allow modeling of the strength of selection by regressing π_N on both viremia and the mutation rate. Under the neutral theory, a negative coefficient for viremia would be expected, indicating the heightened efficacy of purifying selection against deleterious mutations in larger populations. On the other hand, a positive coefficient for the mutation rate would be expected, reflecting the ability of mutation pressure to overcome purifying selection. Other possible developments are technical, including the incorporation of more input formats for SNPGenie, more sophisticated methods accounting for transition/tranvsersion bias, and estimates (*e.g.*, based on Miyata et al. 1979) of chemical distance for nonsynonymous mutations.

Modern evolutionary bioinformatics is a rare discipline in which virtually all analyses of novel data require a combination of substantial amounts of traditional theory in addition to novel input in the form of manual processing, including visual data manipulation and scripting. However, it is important for new tools to combine powerful automation with a degree of flexibility that will also allow their use well into the future. SNPGenie does this by accepting two standard file formats to specify the study genome

and its coding regions—FASTA and GTF (gene transfer format), respectively—but leaving the method for SNP calling relatively open-ended. As such, it is hoped that the software and the extensive labor it represents will prove useful in "making measurable" population genetic parameters, both for increasing our understanding of evolution, and for insights that may help to alleviate the diseases inflicted by pathogens. In finding out the truth, it is my hope that we shall find ourselves immeasurably improved—not just physically, but emotionally and spiritually as well. Lest we get ahead of ourselves, I close with the words of Hughes (1999):

> Finally, it is important to be humble about what we can and cannot know… We must realize that the molecular techniques now available to us have opened a fascinating but limited window on the mechanisms by which over millions of years of [*sic*] life as we know it has evolved. Let us be grateful for that window, while accepting that there will always be much that is mysterious about the history of life on earth.

REFERENCES

Aaskov J, Buzacott K, Thu HM, Lowry K, Holmes EC. 2006. Long-lerm transmission of defective RNA viruses in humans and *Aedes* mosquitoes. Science 311:236–238.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, *et al.* 2015. A global reference for human genetic variation. Nature 526:68–74.

Bächi T. 1988. Direct observation of the budding and fusion of an enveloped virus by video microscopy of viable cells. J. Cell Biol. 107:1689–1695.

Bailey AL, Lauck M, Weiler A, Sibley SD, Dinis JM, Bergman Z, Nelson CW, Correll M, Gleicher M, Hyeroba D, *et al.* 2014. High genetic diversity and adaptive potential of two simian hemorrhagic fever viruses in a wild primate population. PLoS One 9:e90714.

Becker EA, Burns CM, León EJ, Rajabojan S, Friedman R, Friedrich TC, O'Connor SL, Hughes AL. 2012. Experimental analysis of sources of error in evolutionary studies based on Roche/454 pyrosequencing of viral genomes. Genome Biol. Evol. 4:457–465.

Beerenwinkel N, Zagordi O. 2011. Ultra-deep sequencing for the analysis of viral populations. Curr. Opin. Virol. 1:413–418.

Belshaw R, Pybus OG, Rambaut A. 2007. The evolution of genome compression and genomic novelty in RNA viruses. Genome Res. 17:1496–1504.

Bimber BN, Dudley DM, Lauck M, Becker EA, Chin EN, Lank SM, Grunenwald HL, Caruccio NC, Maffitt M, Wilson NA, *et al.* 2010. Whole-genome characterization of human and simian immunodeficiency virus intrahost diversity by ultradeep pyrosequencing. J. Virol. 84:12087–12092.

Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, *et al.* 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816.

Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC. 1987. Structure of the human class I histocompatibility antigen, HLA-A2. Nature 329:506–512.

Brues AM. 1964. The cost of evolution vs. the cost of not evolving. Evolution 18:379–383.

Brüssow H. 2009. The not so universal tree of life or the place of viruses in the living world. Philos. Trans. R. Soc. B 364:2263–2274.

Bull JJ, Sanjuán R, Wilke CO. 2007. Theory of lethal mutagenesis for viruses. J. Virol. 81:2930–2939.

Caì Y, Postnikova EN, Bernbaum JG, Yú S, Mazur S, Deiuliis NM, Radoshitzky SR, Lackemeyer MG, McCluskey A, Robinson PJ, *et al.* 2015. Simian hemorrhagic fever virus cell entry is dependent on CD163 and uses a clathrin-mediated endocytosis-like pathway. J. Virol. 89:844–856.

Callendret B, Bukh J, Eccleston HB, Heksch R, Hasselschwert DL, Purcell RH, Hughes AL, Walker CM. 2011. Transmission of clonal hepatitis C virus genomes reveals the dominant but transitory role of CD8+ T cells in early viral evolution. J. Virol. 85:11833–11845.

Cannon NA, Donlin MJ, Fan X, Aurora R, Tavis JE. 2008. Hepatitis C virus diversity and evolution in the full open-reading frame during antiviral therapy. PLoS One 3:1–12.

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, *et al.* 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat. Genet. 43:956–963.

Chen HY, Di Mascio M, Perelson AS, Ho DD, Zhang L. 2007. Determination of virus burst size in vivo using a single-cycle SIV in rhesus macaques. Proc. Natl. Acad. Sci. USA 104:19079–19084.

Crow JF, Kimura M. 1970. An Introduction to Population Genetics Theory. Caldwell, NJ: The Blackburn Press.

Cullen M, Boland JF, Schiffman M, Zhang X, Wentzensen N, Yang Q, Chen Z, Yu K, Mitchell J, Roberson D, *et al.* 2015. Deep sequencing of HPV16 genomes: A new high-throughput tool for exploring the carcinogenicity and natural history of HPV16 infection. Papillomavirus Res. 1:3–11.

Cutler DJ, Jensen JD. 2010. To pool, or not to pool? Genetics 186:41–43.

de la Torre JC, Holland JJ. 1990. RNA virus quasispecies populations can suppress vastly superior mutant progeny. J. Virol. 64:6278–6281.

den Boer PJ. 1986. The present status of the competitive exclusion principle. Trends Ecol. Evol. 1:25–28.

Dietrich MR. 1994. The origins of the neutral theory of molecular evolution. J. Hist. Biol. 27:21–59.

Doherty PC, Zinkernagel RM. 1975. Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. Nature 256:50–52.

Domingo E. 1992. Genetic variation and quasi-species. Curr. Opin. Genet. Dev. 2:61–63.

Domingo E. 2002. Quasispecies theory in virology. J. Virol. 76:463–465.

Dudley DM, Bailey AL, Mehta SH, Hughes AL, Kirk GD, Westergaard RP, O'Connor DH. 2014. Cross-clade simultaneous HIV drug resistance genotyping for reverse transcriptase, protease, and integrase inhibitor mutations by Illumina MiSeq. Retrovirology 11:122.

Ebrahimi D, Alinejad-Rokny H, Davenport MP. 2014. Insights into the motif preference of APOBEC3 enzymes. PLoS One 9:e87679.

Eigen M, Schuster P. 1977. The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle. Naturwissenschaften 64:541–565.

Eigen M. 1992. Steps Towards Life: A Perspective on Evolution. New York, NY: Oxford University Press.

Eigen M. 1996. On the nature of virus quasispecies. Trends Microbiol. 4:216–218.

Evans DT, O'Connor DH, Jing P, Dzuris JL, Sidney J, da Silva J, Allen TM, Horton H, Venham JE, Rudersdorf RA, *et al.* 1999. Virus-specific cytotoxic T-lymphocyte responses select for amino-acid variation in simian immunodeficiency virus Env and Nef. Nat. Med. 5:1270–1276.

Ewens WJ. 1979. Mathematical Population Genetics. Vol. 9. (Krickeberg K, Levin SA, editors.). Berlin: Springer-Verlag.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. Nat. Rev. Genet. 8:610–618.

Futschik A, Schlötterer C. 2010. The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics 186:207–218.

Galvani AP. 2005. The role of mutation accumulation in HIV progression. Proc. R. Soc. London B 272:1851–1858.

Gellerup DD, Balgeman AJ, Nelson CW, Ericsen AJ, Scarlotta M, Hughes AL, O'Connor SL. 2016. Conditional immune escape during chronic simian immunodeficiency virus infection. J. Virol. 90:545–552.

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M. 2007. What is a gene, post-ENCODE? History and updated definition. Genome Res. 17:669–681.

Godeny EK, de Vries AAF, Wang XC, Smith SL, de Groot RJ. 1998. Identification of the leader-body junctions for the viral subgenomic mRNAs and organization of the simian hemorrhagic fever virus genome: evidence for gene duplication during Arterivirus evolution. J. Virol. 72:862–867.

Graur D, Li W-H. 2000. Fundamentals of Molecular Evolution. Second Edition. Sunderland, MA: Sinauer Associates, Inc., Publishers.

Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E. 2013. On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol. Evol. 5:578–590.

Haldane JBS. 1957. The cost of natural selection. J. Genet. 55:511–524.

Haldane JBS. 1960. More precise expressions for the cost of natural selection. J. Genet. 57:351–360.

Halliburton R. 2004. Introduction to Population Genetics. Upper Saddle River, NJ: Pearson Prentice Hall.

Hanada K, Suzuki Y, Gojobori T. 2004. A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. Mol. Biol. Evol. 21:1074–1080.

Hardin G. 1960. The competitive exclusion principle. Science 131:1292–1297.

Hedskog C, Mild M, Jernberg J, Sherwood E, Bratt G, Leitner T, Lundeberg J, Andersson B, Albert J. 2010. Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. PLoS One 5:e11345.

Holmes EC, Moya A. 2002. Is the quasispecies concept relevant to RNA viruses? J. Virol. 76:460–462.

Holmes EC. 2009. The Evolution and Emergence of RNA Viruses. New York, NY: Oxford University Press.

Howard JC. 1991. Disease and evolution. Nature 352:565–567.

Hughes AL. 1999. Adaptive Evolution of Genes and Genomes. New York, NY: Oxford University Press.

Hughes AL. 2007a. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. Heredity 99:364–373.

Hughes AL. 2007b. Micro-scale signature of purifying selection in Marburg virus genomes. Gene 392:266–272.

Hughes AL. 2008. Near-neutrality: the leading edge of the neutral theory of molecular evolution. Ann. N. Y. Acad. Sci. 1133:162–179.

Hughes AL. 2009. Relaxation of purifying selection on live attenuated vaccine strains of the family Paramyxoviridae. Vaccine 27:1685–1690.

Hughes AL. 2011. Natural selection and the genome. In: Elnitski L, Piontkivska H, Welch LR, editors. Science, Engineering, and Biology Informatics – Vol. 7: Advances in Genomic Sequence Analysis and Pattern Discovery. World Scientific. p. 209–221.

Hughes AL. 2012. Evolution of adaptive phenotypic traits without positive Darwinian selection. Heredity 108:347–353.

Hughes AL, Hughes MAK. 2005. Patterns of nucleotide difference in overlapping and non-overlapping reading frames of papillomavirus genomes. Virus Res. 113:81–88.

Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167–170.

Hughes AL, Yeager M. 1998. Natural selection at major histocompatibility complex loci of vertebrates. Annu. Rev. Genet. 32:415–435.

Hughes AL, Ota T, Nei M. 1990. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Mol. Biol. Evol. 7:515–524.

Hughes AL, Westover K, da Silva J, O'Connor DH, Watkins DI. 2001. Simultaneous positive and purifying selection on overlapping reading frames of the tat and vpr genes of simian immunodeficiency virus. J. Virol. 75:7966–7972.

Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, Yeager M. 2003. Widespread purifying selection at polymorphic sites in human protein-coding loci. Proc. Natl. Acad. Sci. U. S. A. 100:15754–15757.

Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, Yeager M. 2005. Effects of natural selection on interpopulation divergence at polymorphic sites in human protein-coding loci. Genetics 170:1181–1187.

Hughes AL, Piontkivska H, Krebs KC, O'Connor DH, Watkins DI. 2005. Within-host evolution of CD8+-TL epitopes encoded by overlapping and non-overlapping reading frames of simian immunodeficiency virus. Bioinformatics 21:39–44.

Hughes AL, O'Connor S, Dudley DM, Burwitz BJ, Bimber BN, O'Connor D. 2010. Dynamics of haplotype frequency change in a CD8+TL epitope of simian immunodeficiency virus. Infect. Genet. Evol. 10:555–560.

Hughes AL, Becker EA, Lauck M, Karl JA, Braasch AT, O'Connor DH, O'Connor SL. 2012. SIV genome-wide pyrosequencing provides a comprehensive and unbiased view of variation within and outside CD8 T lymphocyte epitopes. PLoS One 7:e47818.

Hunt R, Sauna ZE, Ambudkar S V., Gottesman MM, Kimchi-Sarfaty C. 2009. Silent (synonymous) SNPs: should we care about them? Methods Mol. Biol. 578:23–39.

Ingman M, Gyllensten U. 2009. SNP frequency estimation using massively parallel sequencing of pooled DNA. Eur. J. Hum. Genet. 17:383–386.

Jenkins GM, Worobey M, Woelk CH, Holmes EC. 2001. Evidence for the non-quasispecies evolution of RNA viruses. Mol. Biol. Evol. 18:987–994.

Kew OM, Sutter RW, de Gourville EM, Dowdle WR, Pallansch MA. 2005. Vaccine-derived polioviruses and the endgame strategy for global polio eradication. Annu. Rev. Microbiol. 59:587–635.

Kimchi-Sarfaty C, Oh JM, Kim I-W, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM. 2016. A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science 315:525–528.

Kimura M. 1968. Evolutionary rate at the molecular level. Nature 217:624–626.

Kimura M. 1979. The neutral theory of molecular evolution. Sci. Am. 241:98–126.

Kimura M. 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press.

Klein J, Figueroa F. 1986. Evolution of the major histocompatibility complex. Crit. Rev. Immunol. 6:295–386.

Klein J. 1978. H-2 mutations: their genetics and effect on immune functions. Adv. Immunol. 26:55–146.

Klein J. 1986. Natural History of the Major Histocompatibility Complex. New York, NY: Wiley and Sons.

Knapp EW, Hughes AL. 2012. PolyAna: analyzing synonymous and nonsynonymous polymorphic sites. Conserv. Genet. Resour. 3:429–431.

Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, Kosiol C, Schlötterer C. 2011. Popoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS One 6:e15925.

Kofler R, Pandey RV, Schlötterer C. 2011. PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics 27:3435–3436.

Komar AA. 2007. SNPs, silent but not invisible. Science 315:466–467.

Koonin E V. 2009. Darwinian evolution in the light of genomics. Nucleic Acids Res. 37:1011–1034.

Lauck M, Hyeroba D, Tumukunde A, Weny G, Lank SM, Chapman CA, O'Connor DH, Friedrich TC, Goldberg TL. 2011. Novel, divergent simian hemorrhagic fever viruses in a wild ugandan red colobus monkey discovered using direct pyrosequencing. PLoS One 6:e19056.

Lauck M, Alvarado-Mora M V., Becker EA, Bhattacharya D, Striker R, Hughes AL, Carrilho FJ, O'Connor DH, Pinho JRR. 2012. Analysis of hepatitis C virus intrahost diversity across the coding region by ultradeep pyrosequencing. J. Virol. 86:3952–3960.

Lauck M, Sibley SD, Hyeroba D, Tumukunde A, Weny G, Chapman CA, Ting N, Switzer WW, Kuhn JH, Friedrich TC, *et al.* 2013. Exceptional simian hemorrhagic fever virus diversity in a wild African primate community. J. Virol. 87:688–691.

Lauring AS, Andino R. 2010. Quasispecies theory and the behavior of RNA viruses. PLoS Pathog. 6:e1001005.

Lawlor DA, Zemmour J, Ennis PD, Parham P. 1990. Evolution of class-I MHC genes and proteins: from natural selection to thymic selection. Annu. Rev. Immunol. 8:23–63.

Le T, Chiarella J, Simen BB, Hanczaruk B, Egholm M, Landry ML, Dieckhaus K, Rosen MI, Kozal MJ. 2009. Low-abundance HIV drug-resistant viral variants in treatment-experienced persons correlate with historical antiretroviral use. PLoS One 4:e6079.

Li H, Hughes AL, Bano N, McArdle S, Livingston S, Deubner H, McMahon BJ, Townshend-Bulson L, McMahan R, Rosen HR, *et al.* 2011. Genetic diversity of near genome-wide hepatitis C virus sequences during chronic infection: evidence for protein structural conservation over time. PLoS One 6:e19562.

Li W-H. 1997. Molecular Evolution. Sunderland, MA: Sinauer Associates, Inc., Publishers.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, *et al.* 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478:476–482.

Llewellyn ZN, Salman MD, Pauszek S, Rodriguez LL. 2002. Growth and molecular evolution of vesicular stomatitis serotype New Jersey in cells derived from its natural insect-host: evidence for natural adaptation. Virus Res. 89:65–73.

Lynch M. 2007a. The Origins of Genome Architecture. Sunderland, MA: Sinauer Associates, Inc. Publishers.

Lynch M. 2007b. The frailty of adaptive hypotheses for the origins of organismal complexity. Proc. Natl. Acad. Sci. USA 104:8597–8604.

Lynch M, Bost D, Wilson S, Maruki T, Harrison S. 2014. Population-genetic inference from pooled-sequencing data. Genome Biol. Evol. 6:1210–1218.

Mardis ER. 2008. Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet. 9:387–402.

Martin JA, Wang Z. 2011. Next-generation transcriptome assembly. Nat. Rev. Genet. 12:671–682.

Metzker ML. 2010. Sequencing technologies — the next generation. Nat. Rev. Genet. 11:31–46.

Mishra N, Mathapati BS, Rajukumar K, Nema RK, Behera SP, Dubey SC. 2010. Molecular characterization of RNA and protein synthesis during a one-step growth curve of bovine viral diarrhoea virus in ovine (SFT-R) cells. Res. Vet. Sci. 89:130–132.

Miyata T, Miyazawa S, Yasunaga T. 1979. Two types of amino acid substitutions in protein evolution. J. Mol. Evol. 12:219–236.

Moncla LH, Zhong G, Nelson CW, Dinis JM, Mutschler J, Hughes AL, Watanabe T, Kawaoka Y, Friedrich TC. 2016. Selective bottlenecks shape evolutionary pathways taken during mammalian adaptation of a 1918-like avian influenza virus. Cell Host Microbe 19:169–180.

Moya A, Elena SF, Bracho A, Miralles R, Barrio E. 2000. The evolution of RNA viruses: a population genetics view. Proc. Natl. Acad. Sci. USA 97:6967–6973.

Mudd PA, Ericsen AJ, Burwitz BJ, Wilson NA, O'Connor DH, Hughes AL, Watkins DI. 2012. Escape from CD8(+) T cell responses in Mamu-B*00801(+) macaques differentiates progressors from elite controllers. J. Immunol. 188:3364–3370.

Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE, Gingeras TR, Kent WJ, Birney E, Wold B, *et al.* 2011. A user's guide to the Encyclopedia of DNA elements (ENCODE). PLoS Biol. 9:e1001046.

Nei M. 1975. Molecular Population Genetics and Evolution. In: Neuberger A, Tatum EL, editors. Frontiers of Biology, Volume 40. Vol. 40. Amsterdam: North-Holland Publishing Company. p. 288.

Nei M. 1987. Molecular Evolutionary Genetics. New York, NY: Columbia University Press.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3:418–426.

Nei M, Kumar S. 2000. Molecular Evolution and Phylogenetics. New York, NY: Oxford University Press.

Nei M, Li W-H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76:5269–5273.

Nei M, Suzuki Y, Nozawa M. 2010. The neutral theory of molecular evolution in the genomic era. Annu. Rev. Genomics Hum. Genet. 11:265–289.

Nelson C. 2015. Haldane's dilemma. Inference Int. Rev. Sci. 1:3.

Nelson CW, Hughes AL. 2015. Within-host nucleotide diversity of virus populations: insights from next-generation sequencing. Infect. Genet. Evol. 30:1–7.

Nelson CW, Moncla LH, Hughes AL. 2015. SNPGenie: estimating evolutionary parameters to detect natural selection using pooled next-generation sequencing data. Bioinformatics 31:3709–3711.

Ness RW, Morgan AD, Vasanthakrishnan RB, Colegrave N, Keightley PD. 2015. Extensive de novo mutation rate variation between individuals and across the genome of *Chlamydomonas reinhardtii*. Genome Res. 25:1739–1749.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12:443–451.

Nowak MA, May RM. 2000. Virus Dynamics: Mathematical Principles of Immunology and Virology. New York, NY: Oxford University Press.

O'Connor SL, Becker E a., Weinfurter JT, Chin EN, Budde ML, Gostick E, Correll M, Gleicher M, Hughes AL, Price DA, *et al.* 2012. Conditional CD8+ T Cell Escape during Acute Simian Immunodeficiency Virus Infection. J. Virol. 86:605–609.

Ohta T. 1992. The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. 23:263–286.

Park J-M, Muñoz E, Deem MW. 2010. Quasispecies theory for finite populations. Phys. Rev. E. Stat. Nonlin. Soft Matter Phys. 81:011902.

Pliaka V, Filliponi ME, Kyriakopoulou Z, Ruether IGA, Tsakogiannis D, Gartzonika C, Levidiotou-Stefanou S, Markoulatos P. 2011. Retrospective molecular and phenotypic analysis of poliovirus vaccine strains isolated in Greece. Clin. Microbiol. Infect. 17:1554–1562.

Raineri E, Ferretti L, Esteve-Codina A, Nevado B, Heath S, Pérez-Enciso M. 2012. SNP calling by sequencing pooled samples. BMC Bioinformatics 13:239.

Rammensee H-G, Friede T, Stevanovic S. 1995. MHC ligands and peptide motifs: first lsting. Immunogenetics 41:178–228.

Raoult D, Forterre P. 2008. Redefining viruses: lessons from Mimivirus. Nat. Rev. Microbiol. 6:315–319.

Rellstab C, Zoller S, Tedder A, Gugerli F, Fischer MC. 2013. Validation of SNP allele frequencies determined by pooled next-generation sequencing in natural populations of a non-model plant species. PLoS One 8:e80422.

Rozera G, Abbate I, Bruselles A, Vlassi C, D'Offizi G, Narciso P, Chillemi G, Prosperi M, Ippolito G, Capobianchi MR. 2009. Massively parallel pyrosequencing highlights minority variants in the HIV-1 env quasispecies deriving from lymphomonocyte sub-populations. Retrovirology 6:15.

Sabath N. 2009. Molecular Evolution of Overlapping Genes. Doctoral thesis, University of Houston. ProQuest No. 3405062.

Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. 2010. Viral mutation rates. J. Virol. 84:9733–9748.

Sanjuán R. 2010. Mutational fitness effects in RNA and single-stranded DNA viruses: common patterns revealed by site-directed mutagenesis studies. Philos. Trans. R. Soc. B 365:1975–1982.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. Nat. Biotechnol. 26:1135–1145.

Snijder EJ, Kikkert M, Fang Y. 2013. Arterivirus molecular biology and pathogenesis. J. Gen. Virol. 94:2141–2163.

Strauss JH, Strauss EG. 2008. Viruses and Human Disease. Second Edition. Academic Press.

Stueckemann JA, Ritzi DM, Holth M, Smith MS, Swart WJ, Cafruny WA, Plagemann PGW. 1982. Replication of lactate dehydrogenase-elevating virus in macrophages. 1. Evidence for cytocidal replication. J. Gen. Virol. 59:245–262.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, *et al.* 2015. An integrated map of structural variation in 2,504 human genomes. Nature 526:75–81.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol. Biol. Evol. 30:2725–2729.

Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R. 2006. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature 439:344–348.

Wallace B. 1988. In defense of verbal arguments. Perspect. Biol. Med. 31:201–211.

Wallace B. 1991. Fifty Years of Genetic Load: An Odyssey. Ithaca, NY: Cornell University Press.

Walsh AD, Bimber BN, Das A, Piaskowski SM, Rakasz EG, Bean AT, Mudd PA, Ericsen AJ, Wilson NA, Hughes AL, *et al.* 2013. Acute phase CD8+ T lymphocytes against alternate reading frame epitopes select for rapid viral escape during SIV infection. PLoS One 8:e61383.

Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW. 2007. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. Genome Res. 17:1195–1201.

Wang GP, Sherrill-Mix SA, Chang K-M, Quince C, Bushman FD. 2010. Hepatitis C virus transmission bottlenecks analyzed by deep sequencing. J. Virol. 84:6218–6228.

Watt WB. 2003. Adaptation, constraint, and neutrality: mechanistic case studies with butterflies and their general implications. In: Singh RK, Uyenoyama MK, editors. The Evolution of Population Biology. Cambridge University Press. p. 275–296.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7:256–276.

Wilke CO. 2005. Quasispecies theory in the context of population genetics. BMC Evol. Biol. 5:44.

Wilker PR, Dinis JM, Starrett G, Imai M, Hatta M, Nelson CW, O'Connor DH, Hughes AL, Neumann G, Kawaoka Y, *et al.* 2013. Selection on haemagglutinin imposes a bottleneck during mammalian transmission of reassortant H5N1 influenza viruses. Nat. Commun. 4:2836.

Williamson S, Perry SM, Bustamante CD, Orive ME, Stearns MN, Kelly JK. 2005. A statistical characterization of consistent patterns of human immunodeficiency virus evolution within infected patients. Mol. Biol. Evol. 22:456–468.

Woodruff RC, Huai H, Thompson Jr. JN. 1996. Clusters of identical new mutation in the evolutionary landscape. Genetica 98:149–160.

Woodruff RC, Thompson JN, Gu S. 2004. Premeiotic clusters of mutation and the cost of natural selection. J. Hered. 95:277–283.

Wright CF, Morelli MJ, Thébaud G, Knowles NJ, Herzyk P, Paton DJ, Haydon DT, King DP. 2011. Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. J. Virol. 85:2266–2275.

Wright S. 1931. Evolution in Mendelian populations. Genetics 16:97–159.

Zinkernagel RM, Doherty PC. 1974. Immunological surveillance against altered self components by sensitised T lymphocytes in lymphocytic choriomeningitis. Nature 251:547–548.

APPENDIX A – RIGHTS OF USE

Rights agreement for Chapter 2:

ELSEVIER LICENSE
TERMS AND CONDITIONS

May 09, 2016

This is a License Agreement between Chase Nelson ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier	Elsevier Limited The Boulevard,Langford Lane Kidlington,Oxford,OX5 1GB,UK
Registered Company Number	1982084
Customer name	Chase Nelson
Customer address	2002 Greene St COLUMBIA, SC 29205
License number	3762110959986
License date	Dec 04, 2015
Licensed content publisher	Elsevier
Licensed content publication	Infection, Genetics and Evolution
Licensed content title	Within-host nucleotide diversity of virus populations: Insights from next-generation sequencing
Licensed content author	Chase W. Nelson,Austin L. Hughes
Licensed content date	March 2015
Licensed content volume number	30
Licensed content issue number	n/a
Number of pages	7
Start Page	1
End Page	7
Type of Use	reuse in a thesis/dissertation
Portion	full article
Format	both print and electronic
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Title of your thesis/dissertation	Integrating bioinformatics and geographic information systems to extract evolutionary information from pooled next-generation sequencing variant data

Rights agreement for Chapter 3:

OXFORD UNIVERSITY PRESS LICENSE TERMS AND CONDITIONS

May 09, 2016

This is a License Agreement between Chase Nelson ("You") and Oxford University Press ("Oxford University Press") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Oxford University Press, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3762111148051
License date	Dec 04, 2015
Licensed content publisher	Oxford University Press
Licensed content publication	Bioinformatics
Licensed content title	SNPGenie: estimating evolutionary parameters to detect natural selection using pooled next-generation sequencing data:
Licensed content author	Chase W. Nelson, Louise H. Moncla, Austin L. Hughes
Licensed content date	11/15/2015
Type of Use	Thesis/Dissertation
Institution name	None
Title of your work	Integrating bioinformatics and geographic information systems to extract evolutionary information from pooled next-generation sequencing variant data
Publisher of your work	n/a
Expected publication date	Aug 2016
Permissions cost	0.00 USD
Value added tax	0.00 USD
Total	0.00 USD
Total	0.00 USD
Terms and Conditions	

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.
2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.
3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.
4. No alteration, omission or addition is made to the material without our written consent.