IMPACT OF CYTOSINE INSTABILITY ON SINGLE-STRANDED DNA VIRUS EVOLUTION

By

DANIEL STERN CARDINALE

A dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

And

The Graduate School of Biomedical Sciences

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Microbiology and Molecular Genetics

Written under the direction of

Siobain Duffy

And approved by

____

____

____

____

New Brunswick, New Jersey

October, 2015

ABSTRACT OF THE DISSERTATION

IMPACT OF CYTOSINE INSTABILITY ON SINGLE-STRANDED DNA VIRUS EVOLUTION

By DANIEL STERN CARDINALE

Dissertation Director:
Siobain Duffy

Single-stranded DNA (ssDNA) viruses mutate and evolve at similar rates to RNA viruses, and more rapidly than double-stranded DNA viruses. Unlike RNA viruses, the mechanism underlying these rates is unknown. When unpaired in ssDNA, cytosine is inherently unstable, readily deaminating to uracil. These spontaneous events result in cytosine-to-thymine substitutions, and may explain the high mutation and evolution rates of ssDNA viruses. We examined the codon usage bias of ssDNA bacteriophages and eukaryotic viruses, and found that ssDNA viruses consistently overuse thymine at synonymous sites, regardless of the codon preferences of their hosts. This finding is consistent with a persistent cytosine-to-thymine mutation pressure. We further utilized bottleneck passaging to characterize the mutation spectrum of phiX174, a ssDNA bacteriophage, though we were unable to observe mutational bias. Finally, we treated populations of phiX174 with sodium bisulfite, a cytosine-specific deaminating agent, to induce lethal mutagenesis by elevating the mutation rate of cytosine.

We were able to successfully drive these populations to extinction, and lethal mutagenesis is the most likely explanation for these observations.

Dedication

To Abby, from who I have learned more than any degree could encompass.

Acknowledgements

Rutgers Division of Life Science, Office of Undergraduate Instruction

Rutgers School of Environmental and Biological Sciences

New Jersey Agricultural Experiment Station

Gates Foundation Grant BMGF/UK Department of International Development Phase II OPP105239 “Disease Diagnostics for Sustainable Cassava Production in Africa”

Rutgers Aresty Research Center

Siobain Duffy

Kate DeRosa

Victoria Buchholz

Christopher Lin

Lillian Nichols

Natasia Jacko

Chapter 1 previously published:

Cardinale, D.J., Duffy, S., 2011. Single-stranded genomic architecture constrains optimal codon usage. Bacteriophage 1, 219-224.

Chapter 2 previously published:

Cardinale, D., DeRosa, K., Duffy, S., 2013. Base Composition and Translational Selection are Insufficient to Explain Codon Usage Bias in Plant Viruses. Viruses 5, 162-181.

Table of Contents

Introduction

ssDNA Viruses

Viruses are subcellular, intracellular parasites, biological entities smaller than most cells, and unable to replicate independently from their hosts. They require some part of their host cells for completion of their replication cycles. Whether this means that viruses are not truly alive is a debate that is beyond the scope of this work, but, regardless, they are biologically relevant agents that replicate and evolve using the same mechanisms as cellular life. They are also the most numerous biological entities on earth, and for this reason alone a comprehensive understanding of viral evolution is needed.

Viruses show extreme variation in size, structure, genome composition, replication cycle, and effects on host. These differences make broad classifications difficult, so viruses are most broadly grouped according to their genomic architecture, a scheme called the Baltimore system, after its proposer, David Baltimore (Baltimore, 1971). Genomic architecture refers to the physical makeup of the genome, which dictates the path through which messenger RNA (mRNA) is synthesized. The Baltimore classification system contains the following groups: Group I, double-stranded DNA (dsDNA) viruses, viruses with double-stranded DNA genomes; Group II: single-stranded DNA (ssDNA), viruses with single-stranded DNA genomes; Group III: double-stranded RNA (dsRNA), viruses with double-stranded RNA genomes; Group IV: positive-sense single-stranded RNA ((+)ssRNA), viruses with positive-sense single-stranded RNA genomes; Group V: negative-sense ssRNA ((-)ssRNA), viruses with negative-sense single-stranded RNA genomes; Group VI: reverse-transcribing positive-sense ssRNA (ssRNA-RT), viruses with a positive-sense single-stranded RNA genome that use replicate through a DNA intermediate; and Group VII: reverse-transcribing dsDNA (dsDNA-RT), viruses with dsDNA genomes that replicate through an RNA intermediate.

This work focuses on viruses with ssDNA genomes (Group II). ssDNA genomes tend to be extremely small, ranging in size from less than 1.7 kb, or 1,700 nucleotide bases (Kraberger et al., 2015) to a recently discovered archaeal virus with a 24 kb genome (Mochizukia et al., 2012). The smallest of these genomes encode just two genes: one for the capsid protein that makes up the viral capsid and a second for the replication-associated protein, or Rep/RP, which is required for viral replication, but it is not a DNA polymerase. No known ssDNA viruses encode their own; all are completely dependent on their hosts for genome replication.

These small genomes are extremely compact, with few, short intergenic regions. For instance, the genome of ssDNA bacteriophage phiX174 is 94% coding (Sanger et al., 1978). Many ssDNA viral genomes exhibit overlapping reading frames, which a single stretch of the genome encodes more than a single gene (Barrell et al., 1978). These overlapping genes can occur in the same reading frame, in which the two genes partially overlap, or one is a shortened version of the other. They can also occur in offset reading frames, in which bases in the first codon position of one open reading frame (ORF) are in the second or third positions of the overlapping ORF. Both types of overlapping reading frames are found in the *Escherichia coli* phage phiX174 genome, where gene A* is a shortened version of gene A, while gene B overlaps the A/A* ORF in an alternate reading frame (Barrell et al., 1978).

These small, single-stranded genomes might be expected to show extensive secondary structure caused by intramolecular base pairing, like ssRNA viruses. Single-stranded RNA genomes often utilize complex secondary structures involved in replication, translation, and gene regulation (Andino et al., 1990; Tsukiyama-Kohara et al., 1992; Watts et al., 2009). To the extent that ssDNA genomes exhibit secondary structures, they tend to be a small stem-loops at the site of replication initiation (Boevink et al., 1995; Harding et al., 1993). ssDNA viral genomes also tend to be unordered when encapsidated (Benevides et al., 1991; Incardona et al., 1987; Welsh et al., 1998; Wen et al., 1999), though more recent computational analysis suggests some ssDNA viruses may have persistent and extensive secondary structure (Muhire et al., 2014).

One of the most important features of ssDNA viruses is that they evolve extremely rapidly, as measured by substitution rate, the number of base substitutions per site per year. Higher substitution rates indicate more rapid evolution, while lower substitution rates indicate slower evolutionary change. The fastest-evolving organisms are RNA viruses, which have exhibited substitution rates of 10^{-2}-10^{-5} substitutions/site/year, s/s/yr (Hicks and Duffy, 2014). dsDNA viruses evolve significantly more slowly, approximately 10^{-7}-10^{-9} s/s/yr (Bernard, 1994; Hatwell and Sharp, 2000; McGeoch and Gatherer, 2005).

Because they have a DNA genome and replicate with high-fidelity host polymerases, ssDNA viruses might be expected to evolve at the slower rates of dsDNA viruses. However, ssDNA viruses evolve at rates comparable to RNA viruses (Duffy and Holmes, 2008; Shackelton and Holmes, 2006). It is not known why ssDNA viruses evolve as quickly as RNA viruses, but there are three possible mechanisms: sustained positive selection, shorter generation times, and RNA-like mutation rates.

It is possible that the ssDNA viruses tend to exhibit high substitution rates because those that have been surveyed are often emerging in new hosts, and are therefore experiencing positive selection. In the absence of any selection, substitution rate is a function of mutation rate, but when a virus jumps into a new host, a period of adaption to that host should follow, and this should be reflected in the substitution rate as nonsynonymous changes accumulate. This has been observed in viral emergence events (Shackelton et al., 2005), and when HIV infects a new individual (Ball et al., 2007). Since ssDNA viruses are frequently emergent viruses, we tend to study them when they appear in new hosts, and it is possible that this means that we tend to observe ssDNA viral genomes during periods of rapid adaptive evolution. This positive selection would be reflected in increased fixed, non-synonymous changes in their genomes and therefore elevated substitution rates.

The primary way to measure positive selection is by determining the d_N/d_S ratio for a gene or set of genes over time (Nei and Gojobori, 1986). The d_N/d_S ratio is the ratio of the rate of nonsynonymous substitution to the rate of synonymous substitution in a protein-coding gene. If the d_N/d_S ratio is greater than 1, that is indicative of a larger number of nonsynonymous change, i.e. nucleotide changes that result in a change in amino acid sequence, relative to synonymous changes, i.e. nucleotide changes that do not affect amino acid sequence. More fixed changes to amino acid sequence is indicative of positive selection for those changes, or adaptive evolution, and because more changes will persist and become fixed when they are generally under positive selection, this mechanism can contribute to an elevated nucleotide substitution rate. So during viral emergence, if genomes are experiencing positive selection as the virus adapts to a new host environment, we expect to see a high d_N/d_S ratio, accompanied by higher substitution rates. In these instances, more changes may persist in ssDNA genomes, though the overall number of mutations may not be significantly different from dsDNA genomes. If selection is essentially neutral, synonymous and nonsynonymous changes should accumulate at approximately the same rate, so a d_N/d_S ratio of approximately 1 should be observed. In this situation, selection neither favors nor disfavors changes to amino acid sequences compared to mutations that do not affect amino acid sequence, so the rate at which both accumulate should be about equal. Finally, if purifying or negative selection is acting on a genome, nonsynonymous changes should be selected out, while synonymous mutations should be tolerated and possibly reach fixation. So under purifying selection, a d_N/d_S ratio of less than one is expected.

Therefore, if adaptive evolution is responsible for the high substitution rates observed in ssDNA viruses, then the d_N/d_S ratios in these viruses should be significantly higher than 1, reflecting that positive selection in the form of non-synonymous changes. In some cases of ssDNA viral emergence, it appears that this is the case. One example is the emergence of Canine Parvovirus 2 (CPV2). CPV2 is a viral infection of domesticated dogs that appeared in the 1970s. Its ancestor is feline panleukopenia virus (FPV), a viral infection of cats. CPV2 first appeared in the mid 1970s, and rapidly spread around the world. By the early 1980s, it was ubiquitous in domesticated dogs on every inhabited continent (Parrish and Kawaoka, 2005). During this time, genomic analysis revealed a substitution rate of about 10^{-3} substitutions/site/year, more similar to RNA viruses than dsDNA viruses, and but a fairly low d_N/d_S ratio (about 0.38) (Shackelton et al., 2005). In 1978, a second virus of domesticated dogs emerged: canine parvovirus 2a (CPV2a). CPV2a also evolved from FPV and similarly spread around the world within just a few years of its appearance (Parrish and Kawaoka, 2005). However, while CPV2a evolved at comparable rates to CPV2 and other ssDNA viruses, it exhibited a high d_N/d_S ratio, indicative of strong adaptive evolution. These findings indicate that adaptive selection may contribute to high ssDNA substitution rates, but cannot entirely explain them. Other ssDNA viruses have emerged since exhibited relatively low d_N/d_S ratios, strongly refuting the theory that adaptive evolution drives the high substitution rates of ssDNA viruses (Duffy and Holmes, 2008; Duffy and Holmes, 2009; Shackelton and Holmes, 2006; Shackelton et al., 2005).

Another possible explanation for high ssDNA substitution rates could be shorter generation times. Since mutation rate is measured in terms of mutations per site per replication, but substitutions in terms of s/s/yr, high substitution rates can be achieved with a middling mutation rate by having many more generations per year. This could explain how lower per-generation mutation rates can occur in the same viruses as high per-year substitution rates. There are insufficient data to conclusively accept or reject this possibility, but some ssDNA phages do have shorter generation times than RNA phages of the same host, so it is possible that short generation times contribute to the rapid evolution of ssDNA viruses (Duffy et al., 2008).

The most likely explanation for the rapid evolution of ssDNA viruses is that they experience mutation rates much faster than dsDNA viruses, at rates closer to those of RNA viruses, and consequently experience fixation of mutations at rates comparable to RNA viruses. Unfortunately, the only available data on mutation rates for ssDNA are from ssDNA bacteriophages. While these do mutate more rapidly than dsDNA phages, even controlling for host, they do not appear to mutate quite as rapidly as RNA phages of the same host (Cuevas et al., 2009). However, the differences with dsDNA phages are significant, so faster substitution rates in ssDNA viruses may be attributable, at least in part, to RNA-like mutation rates.

There are different ways to measure mutation rate, but the most common metric is mutations per site per round of genome replication: the number of changes per base each time the genome is copied. A problem arises when comparing across viruses that use different mechanisms to copy their genomes within host cells. Some use a stamping machine model, in which a single or very small number of genomes are used as templates, while most newly synthesized genomes are not. An alternate model is a doubling model in which each genome copy, once synthesized, is used a template to generate subsequent copies (Sanjuán et al., 2010). These differences can confound analyses of mutation rate because, even in selection-free systems, it can be difficult to evaluate the step at which a mutation arose. An alternate metric that has been proposed is changes per base per infected host cell, which might reasonably be interpreted as changes per base per generation, where generation is defined as a complete entry-replication-exit cycle within a single host cell (Sanjuán et al., 2010). Importantly, by either metric, ssDNA viruses do experience mutations at higher rates than dsDNA viruses, though not quite as fast as RNA viruses (Duffy et al., 2008; Sanjuán et al., 2010). The mechanism that explains these higher ssDNA viral mutation rates is unclear, but these measured rates are still not high enough to explain how ssDNA virus substitution rates could equal those of RNA viruses.

The mechanism through which RNA viruses experience rapid mutations is well documented. RNA viruses replicate with an RNA-dependent RNA polymerase (RdRp). Almost no RdRps have the capacity to proofread RNA synthesis while RNA nucleotides are being added to a growing complementary RNA strand. Consequently, errors occur at approximately 10^{-4} mutations/base/replication during genome replication of RNA viruses (Steinhauer and Holland, 1986). Some members of order *Nidovirales* are exceptions; their polymerases exhibit form of proofreading, improving replication fidelity and allowing for much larger RNA genomes than are usually found (Eckerle et al., 2007; Nga et al., 2011). All other RNA viruses, however, are at the mercy of their lower-fidelity RNA polymerases.

Polymerase error cannot explain the high mutation rates observed in ssDNA viruses. No known ssDNA viruses code for their own DNA polymerases; all ssDNA viruses replicate using host DNA polymerases, which are extremely high-fidelity due to their proofreading capabilities. Focusing on ssDNA bacteriophages, the DNA Polymerase III contains a number of subunits, one of which is ε. This protein is responsible for the 3’-to-5’ exonuclease activity of the polymerase complex, which allows the complex detect incorrectly inserted bases during replication, stop, reverse, excise the incorrect nucleotide and replace it. Polymerase proofreading can reduce error rates by two to three orders of magnitude compared to polymerases without proofreading capabilities (Fijalkowska et al., 2012; Rock et al., 2015; St Charles et al., 2015; Zahurancik et al., 2014). Therefore, while ssDNA viruses experience RNA-virus like substitution rates, these rates are achieved through different mechanisms. Specifically, the absence of RdRp proofreading is the cause of high RNA virus mutation rates, while ssDNA viruses utilize their hosts’ high-fidelity DNA polymerases with proofreading capabilities that minimize replication-induced errors. It is unclear how ssDNA viruses would be able to obtain mutation rates as high as RNA viruses, yet they have those similar longer term rates of evolution.

Significance of ssDNA Viruses

ssDNA viruses do not tend to get as many headlines as RNA viruses (Ebola, influenza) or retroviruses (HIV), but they are of immense importance nonetheless. ssDNA viruses of livestock, like Porcine Circovirus 2, result in the deaths of millions of individual animals economic losses of approximately in the tens of millions of pounds in the UK alone (Alarcon et al., 2013). More important than the economic costs of ssDNA viruses are the human costs. Some of the most widespread and damaging ssDNA viruses are those that infect cassava. Cassava is an important source of carbohydrates in much of the world, especially in Africa, so outbreaks can exact extremely high tolls. For example, a 1997 outbreak of a novel recombinant strain of East African Cassava Mosaic Virus in Uganda essentially eradicated that year's cassava crop, resulting in a severe famine (Legg and Thresh, 2000).

Despite the high costs associated with ssDNA virus infection of crops and livestock, there are no cures or treatments available. There are several ways to approach the problem of viral infections. The most attractive option is vaccination; preventing infections is preferable to treatment. There are several ssDNA viruses for which vaccines are available, such as Canine Parvovirus 2 (Yule et al., 1997). However, for many animal-infecting ssDNA viruses, vaccines are not available, and some of the most damaging ssDNA viruses are phytopathogens, for which vaccination is not possible.

When infection does happen, there are antivirals available for some viruses that can mitigate the effects of the infection. One of the most widely used of these is Tamiflu, an anti-influenza compound that competitively inhibits binding between the influenza virus and host cell membrane by binding with the receptors on the surface of the viral particles (Ward, 2005). Other examples include the various drugs used to keep HIV infections at bay. These can be integrase inhibitors, protease inhibitors, or one of a host of other specific drugs that disrupt some part of the HIV life cycle. Unfortunately, no antivirals exist for ssDNA viruses.

One specific subset of antivirals is mutagens, which are used against some RNA viruses. This is possible for two reasons. First, RNA viruses mutate extremely rapidly, so it’s thought that they can be defeated by increasing their mutation rate beyond a certain threshold (Graci and Cameron, 2008). Second, because the mutagens used are RNA base analogs, so they are specific to RNA, they do not adversely affect the viral hosts’ DNA genomes, allowing them to be used without fear of mutagenic damage to the host as well as the virus. This treatment approach has not been attempted in ssDNA viruses, so while there are a number of prevention and/or treatment options available for viruses in general, there are few to none available for ssDNA viruses.

Symptoms of ssDNA viral infection can be treated, but the outcomes are either that the infection self-resolves, or death. In the case of crops infected by ssDNA viruses, the only option is to remove infected plants, so as to prevent the spread of the virus throughout the population, or to simply clear the entire field and start fresh.

Fortunately, there is only one known ssDNA virus that causes disease in humans. It is erythrovirus B19, which is responsible for slapped cheek rash, also known as Fifth Disease. This virus usually infects individuals between the ages of three and fifteen, and by adulthood, over half the population is estimated to have been infected. Symptoms are generally mild in children. They are cold-like symptoms followed by a mild rash, at which point the individual is not longer contagious. The rash can often last several weeks, but more serious symptoms are uncommon (Servey et al., 2007). In adults, B19 infections can have more serious consequences, especially in pregnant women, where miscarriage can result. Adults may also experience arthritis-like joint pain and stiffness, and infections can be serious in immunocompromised patients. Overall, though, B19 is not a serious threat to human survival or well-being (Servey et al., 2007).

That being said, the threat posed by ssDNA viruses should not be underestimated. ssDNA viruses of non-human animals, such as porcine circovirus and chicken anemia virus, have extremely severe symptoms, and it is not only possible, but highly probable that a highly virulent ssDNA virus will one day infect humans. For this reason, the study of ssDNA viruses, specifically the exploration of treatment options, is of paramount importance.

Mutation

Mutations, any changes to the DNA sequence of an organism, are one of the driving forces of evolutionary change. Genetic variation must exist among individuals in order for natural selection to operate, and most of that variation is the result of mutations. Mutations that affect a single base are called point mutations. There are many other types of mutations, which vary in scale from single-base insertions and deletions to large chromosomal inversions and translocations, but single-base mutations are the focus of this work.

The effects of mutations, like the scale, can be extremely diverse. Many mutations are “silent,” meaning that they do not have any apparent effects. In some organisms, such as humans, these are thought to be extremely common, since the vast majority of the human genome is non-coding (Venter et al., 2001). However, in smaller, more dense genomes, mutations to non-coding areas are less frequent, since those regions are a very small percentage of the total genome. While mutations with significant effects can occur outside of protein-coding regions, for example in the upstream regulatory regions that control expression levels, most mutations in viruses with fitness effects can be found in protein-coding regions. Within coding regions, synonymous mutations are those point mutations, usually in the third base of a codon, that preserve the corresponding amino acid sequence despite changing the DNA sequence. These may have few detectable effects. Strictly speaking, synonymous mutations are probably not entire silent; there is probably some degree of selection acting on codon bias (Chamary et al., 2006; Parmley and Hurst, 2007), especially in highly expressed genes (Hockenberry et al., 2014). In general, the fitness effects of synonymous mutations are extremely low compared to those of nonsynonymous mutations.

Nonsynonymous, or missense, mutations are those that affect the amino acid sequence of a gene. Proteins tend to have extremely precise three-dimensional conformations that are closely associated with their function, so most changes to their primary structure that affect this shape will most likely diminish the activity of that protein in some way. This decreased activity usually adversely affects the fitness of the organisms in which they occur. However, some changes to the amino acid sequence of a protein increase the activity of that protein, which can be beneficial or detrimental. If a protein has increased or new activity, that may benefit the organism in which the mutation occurs. A recent example of such a mutation is in the VPU protein of HIV-1, which gained an additional function relative to the same protein of SIV, allowing HIV to overcome a unique feature of the human immune system (Neil et al., 2008).

Mutations can also have the effect of “undoing” previous mutations. Back mutations, also called reversions, are changes that occur at the same site of prior mutations. When a back mutation occurs, the site reverts to the ancestral state, so it appears as though a mutation never occurred. Compensatory mutations are those that occur after a primary mutation has occurred, and result in partial or full recovery of the ancestral phenotype, though not by restoring the ancestral genotype. These mutations are sometimes called second-site mutations. Often, these operate because an initial mutation will result in a change in amino acid sequence that results in altered protein structure, and a second mutation can change an amino acid that interacts with the first, which may result in a structural change back to the ancestral state. In this way, the effects of a compensatory mutation can be similar to those of a back mutation.

The rate at which mutations occur can itself be the product of natural selection. There may be an optimum rate of mutation for many organisms, and a mutation rate too high or too low selected against. For example, phiX174 has a mutation rate of $1.0\text{x}10^{-6}$ mutations/base/replication (Cuevas et al., 2009).

However, it does not utilize one of its host's mutation repair pathways. The tetranucleotide GATC is a recognition site for the methyl-directed mismatch repair system in *E. coli*, which operates through recognition of methylated adenine residues within GATC sites. GATC tetranucleotides occur approximately once every 1,000 bases throughout the *E. coli* genome. Since phiX174 has a 5.4kb genome, it would be expected to contain five to six GATC sites. However, depending on the strain, it contains one or zero. Experimentally modified phiX174 with seven GATC sites experiences a thirty-fold reduction in mutation rate compared to unmodified phiX174 (Cuevas et al., 2011). This is evidence that the mutation rate observed in phiX174 is maintained at a relatively high rate through selection. High mutation rates can also be an active measure taken by pathogens to avoid recognition by the host immune system, and several human pathogens exhibit this trait (Donelson, 1995), though, in general, most mutations are neutral or deleterious (Eyre-Walker and Keightley, 2007).

While mutations are often thought of as random, there are certain bases or sequences that are predisposed to mutate more frequently than others. ssDNA viruses in particular exhibit biased substitution rates (Duffy and Holmes, 2009). Additionally, mutation hot-spots and cold-spots, areas of abnormally high or low mutation frequency, have been observed in many genomes (Chuang and Li, 2004; Galtier, 2006; Paabo, 1996). Repair mechanisms themselves can be biased. In ambiguous cases, where the correct base is no clear, repair mechanisms often preferentially inserts one base, which can result in the net accumulation of that base over time (Galtier et al., 2001).

Mutations occur through a number of different mechanisms, one of which is polymerase error. The enzymes that are responsible for copying DNA or RNA are not perfect; they occasionally insert an incorrect base into a newly synthesized complementary strand. During DNA replication, many of these errors are corrected (Fijalkowska et al., 2012; Pham et al., 1998; Rock et al., 2015; St Charles et al., 2015; Zahurancik et al., 2014), and there are often post-replication correction mechanisms as well.

Spontaneous chemical degradation of bases can also result in changes. Base oxidation is relatively common, and can lead to abnormal bases in DNA, such as 8-oxo-guanine, which can base-pair with adenine, leading to G→T transversions if the oxidized base is not repaired. Spontaneous oxidative deamination can also affect several bases, most notably cytosine, which rapidly experiences spontaneous deamination to uracil, or thymine if methylated. Deamination is the removal of an amine from a molecule, which in this case is replaced with a carbonyl group. Cytosine is particularly susceptible to this kind of damage, even under biological conditions. Oxidative species can launch a nucleophilic attack against the 4’ carbon of the pyrimidine ring, to which the amine group is bound. This causes the opening of the double bond between the 3’ N and 4’ C, the release of ammonia, and the formation of a carbonyl in its place. The C→U mutation ultimately leads to C→T transitions if unrepaired. Free radicals or other oxidizing agents can be strongly mutagenic, and there are a number of cellular mechanisms to clear reactive oxygen species and limit the associated damage (Davies, 1995).

While a mutation rate of $1x10^{-6}$ mutations/base/replication is much higher than that of dsDNA viruses, it is still not high enough to explain the RNA virus-like substitution rates observed in emergent ssDNA viruses. One additional source of mutation in ssDNA viruses than would not been observed in carefully controlled mutation rate assays is cytosine instability. Like thymine and uracil, cytosine is a pyrimidine, a single six-member ring with nitrogen atoms at the 1 and 3 positions. Cytosine features a carbonyl bond at position 2 and an amine group at position 4. It base-pairs with guanine in both DNA and RNA, forming three hydrogen bonds, making a GC base pair a “strong” base pair in Watson-Crick base pairing, in contrast with “weak” AT base pairs, which have only two hydrogen bonds (Figure 1).

Importantly, this spontaneous reaction is one hundred times more likely in unpaired cytosine (Frederico et al., 1990). It is possible that the hydrogen bonds with guanine stabilize the amino group and make the 4’ carbon significantly less susceptible to nucleophilic attack, reducing the incidence of spontaneous deamination by two orders of magnitude. Unlike ssRNA viruses, which often exhibit extensive secondary structure, ssDNA viruses often do not (Benevides et al., 1991; Incardona et al., 1987; Welsh et al., 1998; Wen et al., 1999), leaving cytosines more susceptible to spontaneous oxidative deamination.

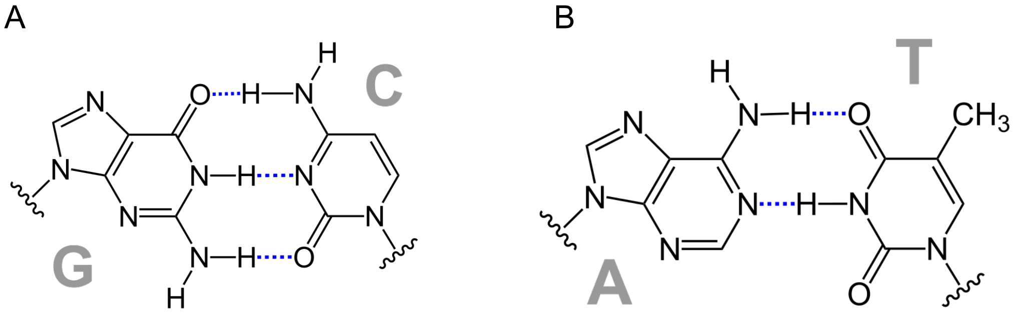


Figure 1. A) GC and B) AT base pairs in DNA. Dotted lines indicate hydrogen bonds. Created by Jü, used under CC0 1.0 public domain dedication.

Bioinformatic evidence suggests that this mechanism is responsible for ssDNA virus mutation rates. A number of ssDNA viruses exhibit a long-term C→T substitution bias, in which the rate of transitions from cytosine to thymine significantly exceeds the rate of the reverse reaction, from thymine to cytosine (Duffy et al., 2008). Spontaneous deamination of cytosine to uracil explains these observations. Many ssDNA viruses with circular genomes replicate via rolling circle replication, in which a single template is used to synthesize a large number of new genomes (Faurez et al., 2009). If cytosine spontaneously changes to uracil through deamination, and the mutation is not corrected, it will be paired with an A in the new DNA strand rather than G during the next round of DNA replication. Every genome synthesized off of that template will insert a T at that site in the newly synthesized genome, rather than a C. The net effect is a transition from C to T. More directly, 5-methylcytosine deaminates directly to thymine without having to go through the uracil intermediate state. Cytosine methylation is a common epigenetic modification (Lister and Ecker, 2009), and is often used by error-correction mechanisms to identify the template strand following DNA replication, so that mispaired bases in the non-methylated newly synthesized strand can be excised and corrected (Marti et al., 2002).

Lethal Mutagenesis

Since mutations are often harmful, they can in theory be used to kill pathogens through lethal mutagenesis. Lethal mutagenesis is increasing the mutation rate within a population sufficiently so that on average, each member of that population produces fewer than one viable offspring. If achieved, this will eventually result in extinction of that population. This mechanism will not kill every individual right away. Rather, it relies on a significant decrease in the average fitness of each member of the population, to the point where the overall reproductive output drops below the level of replacement. When this occurs, the population shrinks, and eventually goes extinct. This can be accomplished by increasing the mutation rate beyond the point at which deleterious mutations can be cleared, so they accumulate in the population over time, decreasing the average fitness of individuals.

How and when lethal mutagenesis occurs depends on a number of factors. The intrinsic mutation rate of the target population in the absence of the mutagen is critical. The closer the population is to the lethal mutagenesis threshold, the easier it will be (requiring a smaller dose of mutagen, for instance) to push them beyond it. The percentage of harmful mutations also plays a role, and has been theorized to be the only factor critical to the success of lethal mutagenesis (Kimura and Maruyama, 1966). Even if many mutations occur, the viruses will not accumulate a significant fitness cost if the mutations are largely neutral or beneficial. Finally, the increase in mutation rate caused by the mutagen will strongly affect dynamics of lethal mutagenesis.

One of the most important aspects of this theory is that a population must exist close to the threshold in order for lethal mutagenesis to be a viable treatment option, and relatively few organisms are thought to fulfill this requirement. RNA viruses, which exhibit the highest viral mutation rates, are thought to be susceptible to lethal mutagenesis, but this strategy has been tried against ssDNA and dsDNA viruses as well (Bull et al., 2013; Domingo-Calap et al., 2012; Paff et al., 2014). Mutagens are also used in other treatment capacities, even if the targets are not necessarily ideal ones for lethal mutagenesis. For example, ethidium bromide is used against some eukaryotic parasites (Roy Chowdhury et al., 2010). Additionally, many anti-cancer chemotherapeutic agents are mutagenic, though are active against disease through other intracellular activities as well.

With the exception of using RNA mutagens against RNA viruses, using mutagens to treat disease posing a significant risk of harm to the host, since the mutagen can affect both the pathogen and host genomes. DNA viruses, bacteria, parasites, and cancer all use DNA as their genetic material, so targeting that for mutagenesis carries a significant risk of harm to the host organisms as well. So mutagenesis is not a commonly used treatment in many cases.

In the case of ssDNA viruses, for which no treatments yet exist, there might be an avenue to use mutagenesis despite their DNA genomes. ssDNA viruses have a long-term cytosine-to-thymine substitution bias (Duffy et al., 2008), which is consistent with an elevated rate of spontaneous mutations from cytosine to uracil (Frederico et al., 1990). This mutation bias may be exploitable since it may be much closer to the lethal mutagenesis threshold than the genome-wide ssDNA virus mutation rate.

There are several potential benefits to this approach. First, unpaired cytosines are significantly more likely, possibly as much as 100 times as likely, to undergo spontaneous deamination than base-paired cytosines (Frederico et al., 1990). It is possible that the hydrogen bonds formed with guanine stabilize the amino group of cytosine. So while a cytosine-specific mutagen might elevate the rate of cytosine deamination in the genomes of ssDNA to dangerously high levels, the more stable, base-paired cytosines of the host will not be affected to such a degree. Furthermore, the cytosine-specific deaminating mutagen sodium bisulfite, which I used in this work, has a higher affinity for ssDNA than for dsDNA. Therefore, it is more likely to affect the ssDNA viral genomes that are already more susceptible to cytosine deamination than the dsDNA genomes of the host organism.

Finally, cellular repair mechanisms can efficiently repair mutations caused by cytosine deamination. Because deamination is such a common occurrence, all cellular life encodes an enzyme that efficiently recognizes uracil in DNA and triggers repair on such mutations (uracil N-glycosylase). Cells also have specific mechanisms to repair methylcytosine deamination to thymine (Kow, 2002). For these reasons, lethal cytosine mutagenesis may be a viable treatment option for ssDNA viruses.

Codon Usage Bias

Patterns of biased mutation show other effects on ssDNA genomes, and one of these may be a signal of C→T mutations in the codon bias of ssDNA viruses. The central dogma of biology is that information stored in DNA is converted to RNA before being translated into proteins. The two main processes of this DNA → RNA → protein information pathway are transcription and translation. Translation is carried out by ribosomes and their associated factors, and results in a protein composed of a specific sequence of amino acids, dictated by the sequence of nucleotides in the messenger RNA. The genetic code is read in groups of three bases, called codons. Because there are four distinct nucleotide bases used in DNA and RNA, there are 64 possible codons. However, only 20 amino acids are used, so there is redundancy in the genetic code. In two cases (methionine and tryptophan), a single codon codes for the amino acid. In the other 18 cases, two, three, four, or six codons code for the same amino acid. Additionally, there are three codons that indicate a stop signal, rather than an amino acid, and these terminate translation.

Different codons that code for the same amino acid are called synonymous (and, as previously mentioned, point mutations that switch among these different codons for the same amino acid are called synonymous mutations). Synonymous codons usually only differ in the third position, also called the wobble position, though there are three amino acids for which there are six synonymous codons, which by necessity vary at the first or second positions as well. Often, a four-fold redundant amino acid will be coded for by all four codons with the same first two bases. For example, proline is coded for the CCN codon family, where N represents any nucleotide. This is possible due to variable base-pairing between the wobble position in the positive sense mRNA and the first base of the tRNA anti-codon. The anticodon often contain irregular or modified bases, such as inosine or 5-oxo-uridine derivatives, which can pair with multiple bases in the wobble position (Agris et al., 2007)

Codon usage bias (CUB) is the unequal usage of synonymous codons within a genome, and is ubiquitous across all forms of life and all types of genomes (Aota and Ikemura, 1986; Shields et al., 1988). No organisms have been documented to have zero codon bias (all synonymous codons used equally), although it is common for some genes within a genome to be more biased than others. Often, many genes have extremely low CUB, or are essentially unbiased, while others, usually expressed at high levels, have very strict codon biases (Bennetzen and Hall, 1982; Hiraoka et al., 2009; Karlin et al., 1998). CUB tends to be consistent within a species. It is uncommon for one gene or set of genes to exhibit one set of codon preferences, while a different gene or set of genes within the same genome exhibits different codon preferences. Typically, some genes are biased, and others are not, but the biased genes tend to prefer the same codons.

Codon preferences between species can vary greatly, and some species have very specific codon preferences. In many cases, the specific reason for these preferences is not clear, though we have a good understanding of why certain codons are over- or under-used in a few cases. For example, human genes tend to use the so-called “rare arginines.” Arginine is one of the six-fold redundant amino acids; the fourfold redundant CGN codons, plus AGR, where R represents purines: adenine and guanine. The two AGR codons for arginine are called the “rare arginine” codons, and are almost never used in *E. coli* (Zhang et al., 1991). In the human genome, however, arginine is coded almost exclusively by the AGR family rather than the otherwise more common CGN family, because CpG dinucleotides are suppressed in the human genome. CpG is recognized by the innate human immune system as foreign, and the presence of CpGs triggers a strong immune response (Greenbaum et al., 2009; Jimenez-Baranda et al., 2011). For this reason, human codons for arginine are almost always from the AGR family rather than the CGN family, unlike many organisms. As a result, human genes expressed in *E. coli* must be modified to match the *E. coli* codon preferences to optimize expression (Burgess-Brown et al., 2008; Kane, 1995).

In most cases of strong codon bias, the underlying reasons are not always clear, though a few trends are common. Codon bias tends to be more pronounced in highly expressed genes compared to genes expressed at lower levels (Hiraoka et al., 2009). A popular hypothesis is that highly biased genes have experienced selection on codon usage to match the frequency of tRNA anticodons, which can facilitate faster elongation of the newly synthesized protein. Highly expressed genes should be under selective pressure for rapid translation. Therefore, it is possible that translational selection, selection on codon usage to optimize the rate of translation by matching tRNA pools, is driving codon bias in highly expressed genes (Hockenberry et al., 2014).

Another potential mechanism that could result in selection for a specific codon bias is selection for translational fidelity. Closely matching codon usage to the tRNA pool contributes to the accuracy of translation (Shah and Gilchrist, 2010; Stoletzki and Eyre-Walker, 2006), and should therefore impose a selection pressure on codon usage, particularly in highly expressed or critical genes. However, there is also evidence that translation proceeding too rapidly can lead to higher inaccuracy in amino acid insertion, so there may be a trade-off involved with translation speed and accuracy (Shah and Gilchrist, 2010).

Finally, codon usage could also affect the speed of translation initiation, though this is less likely to affect codon usage across an entire gene than the other two explanations. The codon bias of the front end of genes, even those with high overall codon bias, tends to be weaker and to the extent that it is present, less similar to the bias in the rest of the gene (Tuller et al., 2010). This may be due to specific codons or bases promoting translation initiation independent of the codons used, creating a “ramp-in” region of each gene. Some data suggest that selection for efficient initiation may overwhelm selection for rapid elongation once translation has begun, which would lead to less or different codon bias near the front of genes (Tuller et al., 2010).

The products of highly expressed genes are more likely to be required in large numbers than those of less expressed genes, and optimal codon use can increase expression levels (Kudla et al., 2009). It is unclear whether this observation this is due in part to accelerated transcription, but by closely correlating a gene to the tRNA pools of a translation system, the translational efficiency of that gene can be improved significantly (Merkl, 2003).

Viruses are dependent on their hosts for translation, so they should experience selection to match the codon biases and tRNA pools of their hosts. Codon bias is a significant factor in phage evolution within bacterial hosts that routinely exhibit codon bias (Carbone, 2008). As expected, phage genomes are under selective pressure imposed by host translational bias, and the most highly affected genes are those coding for highly expressed capsid proteins (Carbone, 2008; Lucks et al., 2008). However, capsid proteins are by no means the only proteins affected by the constraints imposed by host codon bias. Biased genes ranged from about 10% to 80% of all genes in the surveyed genomes, with the average being approximately 30% of genes exhibiting evidence of selection due to host codon bias (Carbone, 2008).

Organisms with high mutational bias, which manifests as extremely high or extremely low GC content, show few of translational selection (Ohama et al., 1990; Wright and Bibb, 1992). When an organism faces very high mutational pressure, it may be unable to adapt its genome to an optimal codon usage pattern, as the mutation rate effectively undoes what selection accomplishes. For example, there is little evidence of consistent codon bias in RNA viruses (Carbone, 2008; Cardinale et al., 2013; Shackelton and Holmes, 2006), which may be attributable to their extremely high mutation rates (Duffy et al., 2008). Mutations may exert such a force on the genomes of RNA phages that selection is unable to keep pace, leading to suboptimal codon usage.

Similarly, if the high C→T substitution rate observed in ssDNA viruses can be explained by rapid cytosine deamination, we expect ssDNA viruses to prefer codons that end in thymine. Over time, as C→T mutations occur, many nonsynonymous changes would be expected to be deleterious, and selected out of ssDNA lineages over time. Conversely, C→T mutations at synonymous sites might be better tolerated, resulting in a preference for T-ending codons, regardless of host codon preferences.

PhiX174: A Model ssDNA Virus

PhiX174 is a ssDNA bacteriophage of family *Microviridae*. It infects *Escherichia coli*, so it is often called a coliphage. Because of its ease of use and safe, abundant host, phiX174 is an extremely common model organism for ssDNA viruses. Its genome was the first DNA genome ever sequenced (Sanger et al., 1977). Its genome is a single, single-stranded, circular DNA molecule, 5,386 bases in length, or 5.386 kb. It contains 11 ORFs: A, A*, B, C, D, E, F, G, H, J, and K (Benbow et al., 1971; Sanger et al., 1978). Like all ssDNA viruses, phiX174 does not encode a polymerase, and must therefore rely upon its host for DNA replication.

Like many ssDNA viral genomes, the phiX174 genome is extremely dense. It is 94% coding, with just 6% composed of short intergenic regions. It also contains extensive overlapping reading frame. Gene A* is encoded entirely within gene A, in the same reading frame, making A* a shortened version of A. Genes A/A* and Gene B partially overlap, and the reading frame of B is offset from the reading frame of A by one base, making the first codon position of each codon in gene B correspond to the third codon position of each overlapping codon in gene A (Weisbeek et al., 1977). Additionally, Gene K is a short ORF that spans the intergenic region between genes A and C, and overlaps both in an alternate reading frame, while gene E is within gene D, also in an alternate reading frame (Sanger et al., 1978).

The icosahedral phiX174 capsid is 33.5 nm in diameter. It is primarily composed of three specific gene products: phiX174 proteins F, G, and H. Sixty copies of Protein F, the major coat protein, make up the primary structure of the capsid. At each of 12 vertices, a spike is present, which interacts with the host cell membrane. It is made of 5 copies of protein G, the major spike protein, and a single H protein, the minor spike protein. Internally, 60 copies of the small protein J, the DNA binding protein, are also present in the mature virion (McKenna et al., 1992)

PhiX174 replicates exclusively using the lytic cycle. Once inside the host cell, it triggers a breakdown of the host genome, using host DNA replication machinery and raw materials to synthesize dozens or hundreds of copies of its own genome. Concurrent with and following DNA replication, viral genes are expressed and viral proteins synthesized, again using host gene expression machinery. Following gene expression, new virions assemble and escape from the host cell, often through virus-induced lysis. PhiX174 protein E mediates host cell lysis though an antibiotic-like mechanism in which the bacterial cell wall is disrupted (Hutchison and Sinsheimer, 1966).

In this thesis, I sought to achieve four specific aims:

1. Determine the relationship between bacteriophage genomic architecture and codon usage bias.

The elevated rate of C→T transitions in ssDNA viruses may constrain their ability to adapt to the codon preferences of their hosts. Because Wan et al. (2004) found that GC content, specifically GC3, can affect codon bias, and because mutations are more likely to be tolerated in the 3rd codon position, single stranded DNA phages were expected to over represent codons ending with thymine, independent of the codon biases exhibited by their hosts.

2. Characterize Geminivirus codon usage bias.

Geminiviruses are a large group of plant-infecting ssDNA viruses with ambisense genomes. They infect both high- and low-GC hosts. Like ssDNA bacteriophages, they were expected to prefer thymine at synonymous sites, regardless of the preferences of their hosts, due to the C to T substitution bias found in ssDNA viruses.

3. Determine the mutation spectrum of a ssDNA bacteriophage.

The mutation and substitution rates of ssDNA viruses have been measured, but the mutation spectrum, the relative rates of possible mutations, has not been characterized. Characterizing the mutation spectrum of a ssDNA bacteriophage will provide a better understanding of ssDNA viral evolution.

4. Use a cytosine specific mutagen to demonstrate the feasibility of lethal mutagenesis against ssDNA viruses.

Since ssDNA viruses experience mutations near the same rate as RNA viruses, they may be susceptible to lethal mutagenesis. To further exploit the tendency of cytosine to deaminate spontaneously, we used a cytosine-specific mutagen to try to demonstrate that by targeting this mutation pressure, populations of ssDNA bacteriophages could be driven to extinction.

References

Agris, P.F., Vendeix, F.A.P., Graham, W.D., 2007. tRNA's Wobble Decoding of the Genome: 40 Years of Modification. J. Mol. Biol. 366, 1-13.

Alarcon, P., Rushton, J., Wieland, B., 2013. Cost of post-weaning multi-systemic wasting syndrome and porcine circovirus type-2 subclinical infection in England – An economic disease model. Prev. Vet. Med. 110, 88-102.

Andino, R., Rieckhof, G.E., Baltimore, D., 1990. A functional ribonucleoprotein complex forms around the 5′ end of poliovirus RNA. Cell 63, 369-380.

Aota, S., Ikemura, T., 1986. Diversity in G+C content at the third position of codons in vertebrate genes and its cause. Nucleic Acids Res. 14, 6345.

Ball, C.L., Gilchrist, M.A., Coombs, D., 2007. Modeling Within-Host Evolution of HIV: Mutation, Competition and Strain Replacement. Bull. Math. Biol. 69, 2361-2385.

Baltimore, D., 1971. Expression of Viral Genomes. Bacteriol. Rev. 35, 235-241.

Barrell, B.G., Shaw, D.C., Walker, J.E., Northrop, F.D., Godson, G.H., Fiddles, J.C., BARCLAY G. BARRELL, D.C.S., * JOHN E. WALKER, FREDERICK D. NORTHROP, GODFREY N. GoDSONt and JOHN C. FIDDES, 1978. Overlapping Genes in Bacteriophages Φx174 and G4. Biochem. Soc. Trans. 6, 63-67.

Benbow, R.M., Hutchison, C.A., Fabricant, J.D., Sinsheimer, R.L., 1971. Genetic Map of Bacteriophage œÜX174. J. Virol. 7, 549-558.

Benevides, J.M., Stow, P.L., Ilag, L.L., Incardona, N.L., Thomas, G.J., 1991. Differences in secondary structure between packaged and unpackaged single-stranded DNA of bacteriophage phi X174 determined by Raman spectroscopy: a model for phi X174 DNA packaging. Biochemistry (Mosc.) 30, 4855-4863.

Bennetzen, J.L., Hall, B.D., 1982. Codon selection in yeast. J. Biol. Chem. 257, 3026-3031.

Bernard, H.-U., 1994. Coevolution of papiliomaviruses with human populations. Trends Microbiol. 2, 140-143.

Boevink, P., Chu, P.W.G., Keese, P., 1995. Sequence of Subterranean Clover Stunt Virus DNA: Affinities with the Geminiviruses. Virology 207.

Bull, J.J., Joyce, P., Gladstone, E., Molineux, I.J., 2013. Empirical Complexities in the Genetic Foundations of Lethal Mutagenesis. Genetics 195, 541-552.

Burgess-Brown, N.A., Sharma, S., Sobott, F., Loenarz, C., Oppermann, U., Gileadi, O., 2008. Codon optimization can improve expression of human genes in Escherichia coli: A multi-gene study. Protein Expr. Purif. 59, 94-102.

Carbone, A., 2008. Codon bias is a major factor explaining phage evolution in translationally biased hosts. J. Mol. Evol. 66, 210-223.

Cardinale, D., DeRosa, K., Duffy, S., 2013. Base Composition and Translational Selection are Insufficient to Explain Codon Usage Bias in Plant Viruses. Viruses 5, 162-181.

Chamary, J.V., Parmley, J.L., Hurst, L.D., 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nature Reviews Genetics 7, 98-108.

Chuang, J.H., Li, H., 2004. Functional Bias and Spatial Organization of Genes in Mutational Hot and Cold Regions in the Human Genome. PLoS Biol. 2, 0253-0263.

Cuevas, J.M., Duffy, S., Sanjuan, R., 2009. Point Mutation Rate of Bacteriophage X174. Genetics 183, 747-749.

Cuevas, J.M., Pereira-Gómez, M., Sanjuán, R., 2011. Mutation rate of bacteriophage ΦX174 modified through changes in GATC sequence context. Infect. Genet. Evol. 11, 1820-1822.

Davies, K.J.A., 1995. Oxidative stress: the paradox of aerobic life. Biochem. Soc. Symp. 61, 1-31.

Domingo-Calap, P., Pereira-Gomez, M., Sanjuan, R., 2012. Nucleoside Analogue Mutagenesis of a Single-Stranded DNA Virus: Evolution and Resistance. J. Virol. 86, 9640-9646.

Donelson, J.E., 1995. Mechanisms of Antigenic Variation in Borrelia hermsii and African Trypanosomes. J. Biol. Chem. 270, 7783-7786.

Duffy, S., Holmes, E.C., 2008. Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. J. Virol. 82, 957-965.

Duffy, S., Holmes, E.C., 2009. Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. J. Gen. Virol. 90, 1539-1547.

Duffy, S., Shackelton, L.A., Holmes, E.C., 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. 9, 267-276.

Eckerle, L.D., Lu, X., Sperry, S.M., Choi, L., Denison, M.R., 2007. High Fidelity of Murine Hepatitis Virus Replication Is Decreased in nsp14 Exoribonuclease Mutants. J. Virol. 81, 12135-12144.

Eyre-Walker, A., Keightley, P.D., 2007. The distribution of fitness effects of new mutations. Nature Reviews Genetics 8, 610-618.

Faurez, F., Dory, D., Grasland, B., Jestin, A., 2009. Replication of porcine circoviruses. Virol. J. 6, 60.

Fijalkowska, I.J., Schaaper, R.M., Jonczyk, P., 2012. DNA replication fidelity in Escherichia coli: a multi-DNA polymerase affair. FEMS Microbiol. Rev. 36, 1105-1121.

Frederico, L.a., Kunkel, T.a., Shaw, B.R., 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry (Mosc.) 29, 2532-2537.

Galtier, N., 2006. Mutation hot spots in mammalian mitochondrial DNA. Genome Res. 16, 215-222.

Galtier, N., Piganeau, G., Mouchiroud, D., Duret, L., 2001. GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. Genetics 159, 907-911.

Graci, J.D., Cameron, C.E., 2008. Therapeutically targeting RNA viruses via lethal mutagenesis. Future Virology 3, 553-566.

Greenbaum, B.D., Rabadan, R., Levine, A.J., 2009. Patterns of Oligonucleotide Sequences in Viral and Host Cell RNA Identify Mediators of the Host Innate Immune System. PLoS ONE 4, e5969.

Harding, R.M., Burns, T.M., Hafner, G., Dietzgen, R.G., Dale, J.L., 1993. Nucleotide sequence of one component of the banana bunchy top virus genome contains a putative replicase gene. J. Gen. Virol. 74, 323-328.

Hatwell, J.N., Sharp, P.M., 2000. Evolution of human polyomavirus JC. J. Gen. Virol. 81, 1191-1200.

Hicks, A.L., Duffy, S., 2014. Cell Tropism Predicts Long-term Nucleotide Substitution Rates of Mammalian RNA Viruses. PLoS Pathog. 10, e1003838.

Hiraoka, Y., Kawamata, K., Haraguchi, T., Chikashige, Y., 2009. Codon usage bias is correlated with gene expression levels in the fission yeast Schizosaccharomyces pombe. Genes Cells 14, 499-509.

Hockenberry, A.J., Sirer, M.I., Amaral, L.A.N., Jewett, M.C., 2014. Quantifying Position-Dependent Codon Usage Bias. Mol. Biol. Evol. 31, 1880-1893.

Hutchison, C.A., Sinsheimer, R.L., 1966. The process of infection with bacteriophage Œ¶X174: X. Mutations in a Œ¶X lysis gene. J. Mol. Biol. 18, 429-IN422.

Incardona, N.L., Prescott, B., Sargent, D., Lamba, O.P., Thomas, G.J., 1987. Phage phi X174 probed by laser Raman spectroscopy: evidence for capsid-imposed constraint on DNA secondary structure. Biochemistry (Mosc.) 26, 1532-1538.

Jimenez-Baranda, S., Greenbaum, B., Manches, O., Handler, J., Rabadan, R., Levine, A., Bhardwaj, N., 2011. Oligonucleotide Motifs That Disappear during the Evolution of Influenza Virus in Humans Increase Alpha Interferon Secretion by Plasmacytoid Dendritic Cells. J. Virol. 85, 3893-3904.

Kane, J.F., 1995. Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coil. Curr. Opin. Biotechnol. 6, 494-500.

Karlin, S., Mrázek, J., Campbell, a.M., 1998. Codon usages in different gene classes of the Escherichia coli genome. Mol. Microbiol. 29, 1341-1355.

Kimura, M., Maruyama, T., 1966. The Mutational Load with Epistatic Gene Interactions in Fitness. Genetics 54, 1337-1351.

Kow, Y.W., 2002. Repair of deaminated bases in DNA12. Free Radic. Biol. Med. 33, 886-893.

Kraberger, S., Argüello-Astorga, G.R., Greenfield, L.G., Galilee, C., Law, D., Martin, D.P., Varsani, A., 2015. Characterisation of a diverse range of circular replication-associated protein encoding DNA viruses recovered from a sewage treatment oxidation pond. Infect. Genet. Evol. 31, 73-86.

Kudla, G., Murray, A.W., Tollervey, D., Plotkin, J.B., 2009. Coding-Sequence Determinants of Gene Expression in Escherichia coli. Science 324, 255-258.

Legg, J.P., Thresh, J.M., 2000. Cassava mosaic virus disease in East Africa: a dynamic disease in a changing environment. Virus Res. 71, 135-149.

Lister, R., Ecker, J.R., 2009. Finding the fifth base: Genome-wide sequencing of cytosine methylation. Genome Res. 19, 959-966.

Lucks, J.B., Nelson, D.R., Kudla, G.R., Plotkin, J.B., 2008. Genome landscapes and bacteriophage codon usage. PLoS Comput. Biol. 4, e1000001.

Marti, T.M., Kunz, C., Fleck, O., 2002. DNA Mismatch Repair and Mutation Avoidance Pathways. J. Cell. Physiol. 191, 28-41.

McGeoch, D.J., Gatherer, D., 2005. Integrating Reptilian Herpesviruses into the Family Herpesviridae. J. Virol. 79, 725-731.

McKenna, R., Xia, D., Willingmann, P., Iiag, L.L., Krishnaswamy, S., Rossmann, M.G., Olson, N.H., Baker, T.S., Incardona, N.L., 1992. Atomic structure of single-stranded DNA bacteriophage ΦX174 and its functional implications. Nature 355, 137-143.

Merkl, R., 2003. A Survey of Codon and Amino Acid Frequency Bias in Microbial Genomes Focusing on Translational Efficiency. J. Mol. Evol. 57, 453-466.

Mochizukia, T., Krupovica, M., Pehau-Arnaudetb, G.r., Sakoc, Y., Forterrea, P., Prangishvilia, D., 2012. Archaeal virus with exceptional virion architecture and the largest single-stranded DNA genome. Proceedings of the National Academy of Sciences 109, 13386-13391.

Muhire, B.M., Golden, M., Murrell, B., Lefeuvre, P., Lett, J.-M., Gray, A., Poon, A.Y.F., Ngandu, N.K., Semegni, Y., Tanov, E.P., Monjane, A.r.L., Harkins, G.W., Varsani, A., Shepherd, D.N., Martina, D.P., 2014. Evidence of Pervasive Biologically Functional Secondary Structures within the Genomes of Eukaryotic Single-Stranded DNA Viruses. J. Virol. 88, 1972-1989.

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

Neil, S.J.D., Zang, T., Bieniasz, P.D., 2008. Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. Nature 451, 425-430.

Nga, P.T., Parquet, M.d.C., Lauber, C., Parida, M., Nabeshima, T., Yu, F., Thuy, N.T., Inoue, S., Ito, T., Okamoto, K., Ichinose, A., Snijder, E.J., Morita, K., Gorbalenya, A.E., 2011. Discovery of the First Insect Nidovirus, a Missing Evolutionary Link in the Emergence of the Largest RNA Virus Genomes. PLoS Pathog. 7, e1002215.

Ohama, T., Muto, a., Osawa, S., 1990. Role of GC-biased mutation pressure on synonymous codon choice in Micrococcus luteus, a bacterium with a high genomic GC-content. Nucleic Acids Res. 18, 1565-1569.

Paabo, S., 1996. Mutational Hot Spots in the Mitochondrial Microcosm. Am. J. Hum. Genet. 59, 493-496.

Paff, M.L., Stolte, S.P., Bull, J.J., 2014. Lethal Mutagenesis Failure May Augment Viral Adaptation. Mol. Biol. Evol. 31, 96-105.

Parmley, J.L., Hurst, L.D., 2007. How do synonymous mutations affect fitness? Bioessays 29, 515-519.

Parrish, C.R., Kawaoka, Y., 2005. THE ORIGINS OF NEW PANDEMIC VIRUSES: The Acquisition of New Host Ranges by Canine Parvovirus and Influenza A Viruses. Annu. Rev. Microbiol. 59, 553-586.

Pham, P.T., Olson, M.W., McHenry, C.S., Schaaper, R.M., 1998. The Base Substitution and Frameshift Fidelity of Escherichia coli DNA Polymerase III Holoenzyme in Vitro. J. Biol. Chem. 273, 23575-23584.

Rock, J.M., Lang, U.F., Chase, M.R., Ford, C.B., Gerrick, E.R., Gawande, R., Coscolla, M., Gagneux, S., Fortune, S.M., Lamers, M.H., 2015. DNA replication fidelity in Mycobacterium tuberculosis is mediated by an ancestral prokaryotic proofreader. Nat. Genet. advance online publication.

Roy Chowdhury, A., Bakshi, R., Wang, J., Yildirir, G., Liu, B., Pappas-Brown, V., Tolun, G., Griffith, J.D., Shapiro, T.A., Jensen, R.E., Englund, P.T., 2010. The Killing of African Trypanosomes by Ethidium Bromide. PLoS Pathog. 6, e1001226.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., III, C.A.H., Slocombe, P.M., Smith, M., 1977. Nucleotide sequence of bacteriophage ΦX174 DNA. Nature 265.

Sanger, F., Coulson, A.R., Friedmann, T., Air, G.M., Barrell, B.B., N. L. Brown, J.C.F., III, C.A.H., Slocombe, P.M., Smith, M., 1978. The Nucleotide Sequence of ΦX174. J. Mol. Biol. 125, 225-246.

Sanjuán, R., Nebot, M.R., Chirico, N., Mansky, L.M., Belshaw, R., 2010. Viral mutation rates. J. Virol. 84, 9733-9748.

Servey, J.T., Reamy, B.V., Hodge, J., 2007. Clinical Presentations of Parvovirus B19 Infection. Am. Fam. Physician 75, 373-376.

Shackelton, L.A., Holmes, E.C., 2006. Phylogenetic evidence for the rapid evolution of human B19 erythrovirus. J. Virol. 80, 3666.

Shackelton, L.a., Parrish, C.R., Truyen, U., Holmes, E.C., 2005. High rate of viral evolution associated with the emergence of carnivore parvovirus. Proc. Natl. Acad. Sci. U. S. A. 102, 379-384.

Shah, P., Gilchrist, M.A., 2010. Effect of Correlated tRNA Abundances on Translation Errors and Evolution of Codon Usage Bias. PLoS Genet. 6, e1001128-e1001128.

Shields, D.C., Sharp, P.M., Higgins, D.G., Wright, F., 1988. "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. 5, 704-716.

St Charles, J.A., Liberti, S.E., Williams, J.S., Lujan, S.A., Kunkel, T.A., 2015. Quantifying the contributions of base selectivity, proofreading and mismatch repair to nuclear DNA replication in Saccharomyces cerevisiae. DNA Repair 31, 41-51.

Steinhauer, D.A., Holland, J.J., 1986. Direct method for quantitation of extreme polymerase error frequencies at selected single base sites in viral RNA. J. Virol. 57, 219-228.

Stoletzki, N., Eyre-Walker, A., 2006. Synonymous Codon Usage in Escherichia coli: Selection for Translational Accuracy. Mol. Biol. Evol. 24, 374-381.

Tsukiyama-Kohara, K., IIzuka, N., Kohara, M., Nomoto, A., 1992. Internal Ribosome Entry Site within Hepatitis C Virus RNA. J. Virol. 66, 1476-1483.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., Pilpel, Y., 2010. An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. Cell 141, 344-354.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V.D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.-R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z.Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S.C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guig√≥, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A.D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., Zhu, X., 2001. The Sequence of the Human Genome. Science 291, 1304-1351.

Ward, P., 2005. Oseltamivir (Tamiflu(R)) and its potential for use in the event of an influenza pandemic. J. Antimicrob. Chemother. 55, i5-i21.

Watts, J.M., Dang, K.K., Gorelick, R.J., Leonard, C.W., Bess Jr, J.W., Swanstrom, R., Burch, C.L., Weeks, K.M., 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. Nature 460, 711-716.

Weisbeek, P.J., Borrias, W.E., Langeveld, S.A., Baas, P.D., Arkel, G.A.V., 1977. Bacteriophage ΦX174: Gene A overlaps gene B. Proc. Natl. Acad. Sci. U. S. A. 74, 2504-2508.

Welsh, L.C., Marvin, D.A., Perham, R.N., 1998. Analysis of X-ray diffraction from fibres of Pf1 Inovirus (filamentous bacteriophage) shows that the DNA in the virion is not highly ordered. J. Mol. Biol. 284, 1265-1271.

Wen, Z.Q., Armstrong, A., Thomas Jr, G.J., 1999. Demonstration by ultraviolet resonance Raman spectroscopy of differences in DNA organization and interactions in filamentous viruses Pf1 and fd. Biochemistry (Mosc.) 38, 3148-3156.

Wright, F., Bibb, M.J., 1992. Codon usage in the G+C-rich Streptomyces genome. Gene 113, 55-65.

Yule, T.D., Roth, M.B., Dreier, K., Johnson, A.F., Palmer-Densmore, M., Simmons, K., Fanton, R., 1997. Canine parvovirus vaccine elicits protection from the inflammatory and clinical consequences of the disease. Vaccine 15, 720-729.

Zahurancik, W.J., Klein, S.J., Suo, Z., 2014. Significant contribution of the 3,Ä≤,Üí5,Ä≤ exonuclease activity to the high fidelity of nucleotide incorporation catalyzed by human DNA polymerase œµ. Nucleic Acids Res. 42, 13853-13860.

Zhang, S., Zubay, G., Goldman, E., 1991. Low-usage codons in Escherichia coli, yeast, fruit fly and primates. Gene 105, 61-72.

Chapter 1

Single-stranded genomic architecture constrains optimal codon usage

Abstract

Viral codon usage is shaped by the conflicting forces of mutational pressure and selection to match host patterns for optimal expression. We examined whether genomic architecture (single- or double-stranded DNA) influences the degree to which bacteriophage codon usage differ from their primary bacterial hosts and each other. While both correlated equally with their hosts' genomic nucleotide content, the coat genes of ssDNA phages were less well adapted than those of dsDNA phages to their hosts' codon usage profiles due to their preference for codons ending in thymine. No specific biases were detected in dsDNA phage genomes. In all nine of ten cases of codon redundancy in which a specific codon was overrepresented, ssDNA phages favored the NNT codon. A cytosine to thymine biased mutational pressure working in conjunction with strong selection against non-synonymous mutations appears be shaping codon usage bias in ssDNA viral genomes.

Introduction

Viruses usually exhibit genomic signatures that closely mimic those of their primary hosts (Carbone, 2003; Sharp and Li, 1987), in part to better evade innate and acquired immune responses (Sharp, 1986; Wong et al., 2010). However, the majority of the close adherence to host nucleotide usage is attributed to selection for improved translational speed and efficiency, which are correlates of viral fitness. Synonymous codons are used at different frequencies in virtually all organisms (Aota and Ikemura, 1986; Shields et al., 1988), and the most frequently used codons correlate with the most abundant tRNAs within a cell (Ikemura, 1981, 1982). These favored synonymous codons are therefore recognized (Curran and Yarus, 1989) and translated (Bennetzen and Hall, 1982; Pedersen, 1984; Robinson et al., 1984) more rapidly. The most frequently expressed cellular genes within a given organism exhibit similar patterns of this codon usage bias (CUB) and are more biased than less frequently expressed genes (Bennetzen and Hall, 1982; Hershberg and Petrov, 2008; Hiraoka et al., 2009; Karlin et al., 1998; Sharp and Li, 1986). For viruses, these factors should contribute to increased rate of replication when strictly adhering to host CUB. Therefore many viruses have been under selective pressure to match the CUB of their preferred hosts (Sharp et al., 1985). Despite increased attention to the genomic match between viruses and their hosts, there have been few studies examining how different viral genomic architectures facilitate or hinder adaptation to their hosts' genomes.

Phages are the optimal system in which to explore how genomic architecture affects viral molecular evolution. The codon bias expressed in prokaryotic hosts is constant for each host cell, unlike multi-cellular organisms, in which codon usage profiles are affected by tissue-specific gene expression (Duret and Mouchiroud, 1999). Perhaps due to this, phage are more strongly adapted to their primary hosts' CUB than eukaryotic viruses (Bahir et al., 2009), allowing the greatest potential to identify factors that diminish the match between virus and host genomes. Bacterial hosts also offer a wider range of genomic nucleotide content to examine compared to plant or mammalian hosts, and their CUB have been well-documented. Additionally, while phage host ranges are far from perfectly annotated, bacteriophage host ranges are usually quite narrow (Hyman and Abedon, 2010) and many of their host ranges have been better delineated than eukaryotic viruses, such as phytopathogens (Bahir et al., 2009) .

Two distinct phage genomic architectures (single-stranded DNA, ssDNA and double-stranded DNA, dsDNA) have been amply sequenced; unfortunately, the small number of sequenced RNA phages precludes their close examination at this time. The two DNA-based architectures are subject to specific constraints: dsDNA phages can house the largest genomes, up to ~300 kb (Serwer et al., 2004; Thomas et al., 2008), whereas even the largest ssDNA phages are smaller than 10 kb (Kawasaki et al., 2007). Many dsDNA phages encode their own tRNAs, (*e.g.*, T4 encodes eight (Miller et al., 2003)), decreasing selection for adherence to host CUB, whereas none have been found in ssDNA phages. dsDNA phages have the lowest mutation rates among viruses, while ssDNA phage mutation rates are faster, approaching those of a dsRNA phage (Drake, 1991; Sanjuán et al., 2010). Eukaryotic viruses with the same ssDNA genomic architecture exhibit evolutionary rates orders of magnitude above those seen in eukaryotic dsDNA viruses (Duffy et al., 2008). Consequently, faster-evolving ssDNA phages might be better able to adapt to host-imposed genomic conditions. Conversely, the mutation frequency in ssDNA phages may diminish their ability to conform to their host codon preferences.

Genomic GC content is a rough predictor of CUB, and many viruses match the GC content of their hosts (Adams and Antoniw, 2004; Antezana and Kreitman, 1999; Bernardi and Bernardi, 1986; Karlin and Mrázek, 1996; Sueoka and Kawanishi, 2000). Bacteriophage GC content, in particular, correlates strongly to that of their primary bacterial hosts (Xia and Yuen, 2005). We measured the similarity in GC content between each ssDNA and dsDNA GenBank phage reference genome and that of its primary host. We used the most numerous group of phages with a common host, *Escherichia coli*, to compare codon adaptation indices (CAI) and relative synonymous codon usage (RSCU) for a subset of highly expressed genes from dsDNA and ssDNA coliphages (Sharp and Li, 1987). Our results show that genomic architecture correlates to statistically significant differences in nucleotide content and codon usage between ssDNA and dsDNA phages, and point to an enrichment of thymine as a cause.

Materials and Methods

All available ssDNA and dsDNA bacteriophage genome reference sequences were collected from GenBank on March 16, 2011. Reference sequences were used to avoid biasing our data sets towards any particular phage species, or highly studied phage, such as the model organisms PhiX174 or T7. These genomes were separated according to genomic architecture for further analysis. Initially collected were 41 ssDNA phages and 447 dsDNA phages. For each phage having a known host with a sequenced genome (GenBank reference sequence), the relationship between the GC content of the phage and the host bacterium was examined. Because not every sequenced phage has an identified and sequenced host, not all phages were included in this analysis. Four ssDNA phages were excluded, as were 44 dsDNA phages.

The codon usage biases of representative ssDNA and dsDNA phages were examined to gain a more complete picture of the CUB patterns in both architectures. Codon usage profiles were determined using major coat/capsid genes, or, in the eight cases for which coat genes were not available, tail gene sequences retrieved from GenBank reference genomes. These structural proteins are highly expressed and exhibit the highest degrees of codon usage bias found in phage.(Carbone, 2008; Lucks et al., 2008) We compared codon usage between the two genomic architectures for phages infecting a single host: *Escherichia coli*. Coat or tail genes from 11 ssDNA and 34 dsDNA phages were used. The online CAIcal tool (Puigbò et al., 2008) was used to calculate each phage's codon adaptation index (CAI), a measure of the degree to which one gene or set of genes adheres to the CUB of another gene or set of genes (Sharp and Li, 1987), as implemented by Xia (Xia, 2007Xia, 2007). CAI ranges from zero to one; values closer to one indicate a strong correlation. The average CAI was calculated for both architectures.

Frequency of GC in the first and second codon positions (GC1,2) and in the third position (GC3) were calculated for these genes using CAIcal and relationship between the two was analyzed. A plot of GC1,2 against GC3 is a common measure of the factors affecting CUB in a gene or set of genes; a strong correlation between the two implies that genome-wide mutational pressures are the driving force behind CUB, while a weaker correlation indicates that some force is unequally affecting the first two positions and the third position. Usually, this is interpreted as implying a selective force acting on CUB, as is expected to be the case for viruses under relatively strong selection for translational speed.

To examine the variation in codon usage that contributes to the differing CAI values and site-specific base compositions, relative synonymous codon usage (RSCU) values were calculated for the same sets of genes using CAIcal. RSCU is a measure of the relative codon usage for each individual degenerate amino acid compared to expected levels if synonymous codons were used with equal frequency. An RSCU of about one indicates that a codon is used as frequently as expected, while values above or below one indicate over or underuse of that synonymous codon, respectively. Mean dsDNA coliphage RSCUs were compared to ssDNA coliphage RSCU to determine the proximate cause of the observed variation in CAI. RSCU was also calculated for 17 additional sufficiently well-annotated genomes of ssDNA phages infecting a wide host range (primarily infecting *Acholplasma, Bdellovibrio*, *Chlamydia*, *Escherichia*, *Propionibacteria*, *Pseudomonas*, *Ralstonia*, *Spiroplasma*, *Vibrio*, and *Xanthomonas*, and the complete set of 28 ssDNA phage RSCUs was assessed for consistent CUB. For amino acids with six-fold redundancy (L, R, S), RSCUs were calculated separately for the codon sets with four-fold and two-fold redundancy. Significantly biased codon use was measured for each codon with one-tailed t-tests (Microsoft Excel) and Bonferroni correction for multiple comparisons (alpha = 0.017 for 4-fold, alpha = 0.025 for 3-fold).

Results and Discussion

GC content in ssDNA and dsDNA phages was highly correlated to host GC content (r^2=0.82 for ssDNA phages, 0.84 for dsDNA phages, equally correlated p=0.72) across a very wide range of host GC content (~0.25 to ~0.72) (Figure 1). A previous study found significant differences between ssDNA and dsDNA phage nucleotide correlation with their hosts,(Xia and Yuen, 2005) but the additional 333 dsDNA and 13 ssDNA reference sequences added to GenBank since that analysis suggest there is no difference (Supplementary Figure 1). ssDNA phages exhibited a pronounced genomic thymine bias (average 0.30 T), but nonetheless infected hosts with a range of GC contents (0.25 to 0.70), as wide as that of dsDNA phages (0.26 to 0.72).

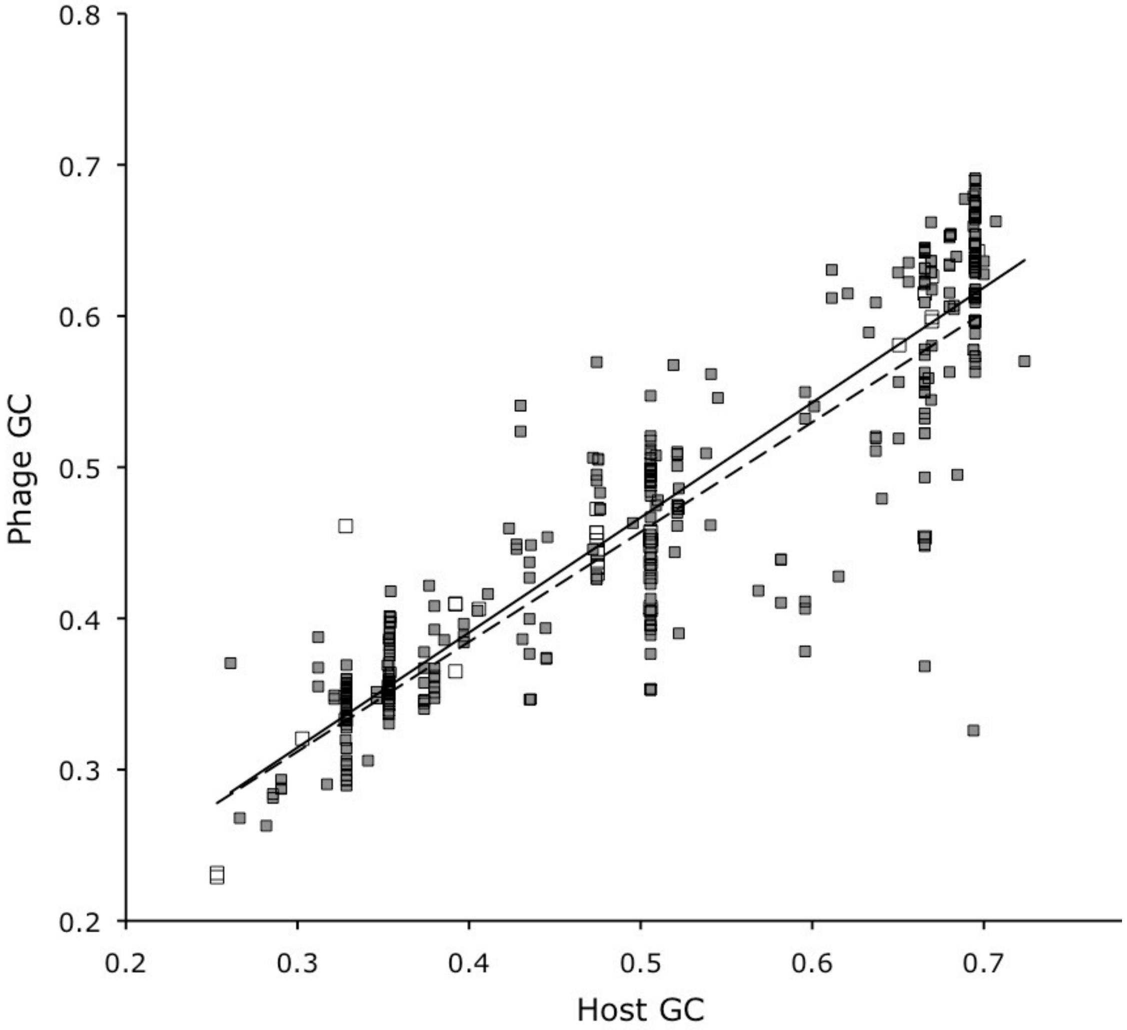


Figure 1. Correlation between host and phage genomic GC content. Grey squares indicate dsDNA phages, open squares ssDNA. Best-fit linear regression lines are solid for dsDNA (r^2=0.84) and dashed for ssDNA (r^2=0.82). There was no significant difference between the correlations (p=0.72).

Correlated GC content was a poor predictor of strong CAI match between *E. coli* and the coat genes of its phages. The mean CAI of ssDNA coliphages was 0.706, while the dsDNA phages were significantly better matched to *E. coli* (0.744, p<0.001, Figure 2). This number includes eight dsDNA coliphage genomes for which tail protein encoding genes were used, rather that coat protein encoding genes, due to the absence of properly annotated coat genes. The inclusion of tail genes did not change the results of this analysis (p<0.001 with and without the eight tail genes). The evidence of selection for translational efficiency is stronger for dsDNA phages.

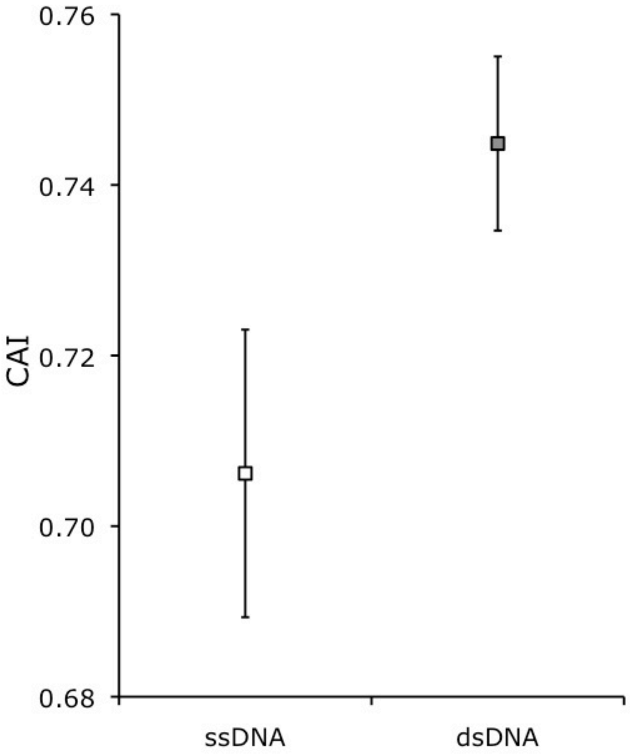


Figure 2. Mean coat gene CAI with 95% confidence intervals of ssDNA (n=11), dsDNA (n=34) coliphages.

Comparison of the GC content of the first two positions of each codon (GC1,2) and the third position (GC3) of these genes revealed an interesting pattern: for both ssDNA and dsDNA coliphages, the GC1,2 was restricted to a tight range between about 0.45 and 0.55. dsDNA GC3 varied along a wide range, from 0.26 to 0.69, but ssDNA GC3 occupied a narrower range, from 0.30 to 0.54 (Figure 3). Furthermore, when plotted with a line representing a perfect correlation between GC1,2 and GC3, all but one of the ssDNA phages fell to the left of that line (Figure 3), indicating a paucity of GC in the third codon position of their coat genes. Conversely, the dsDNA coat genes were GC3-rich or GC3-poor in approximately equal numbers. Past studies have indicated that strong mutational biases often occur with low levels of CUB,(Andersson and Sharp, 1996; Ohama et al., 1990; Ohkubo et al., 1987) possibly because a strong, non-specific mutational pressure would prevent any persistent, directional changes in the genome. The consistently lower GC3 content of the ssDNA genes suggests that a specific mutational pressure might be reducing GC3 content in a directional manner, which is disrupting the effects of selection for translational efficiency.

We further investigated the GC3-poor nature of ssDNA coliphage coat proteins with RSCU analysis. It revealed statistically significant variation in use for 15 of 59 codons between ssDNA and dsDNA phage ($p<0.03$ for TTG, $p<0.002$ for CTT and TCC, $p<0.001$ for all other codons, Figure 4). Notably, for four of the five codons more frequently used by ssDNA rather than dsDNA coliphages, thymine was in the third position. No codons enriched in dsDNA phage relative to ssDNA phage contained thymine in the third positions.

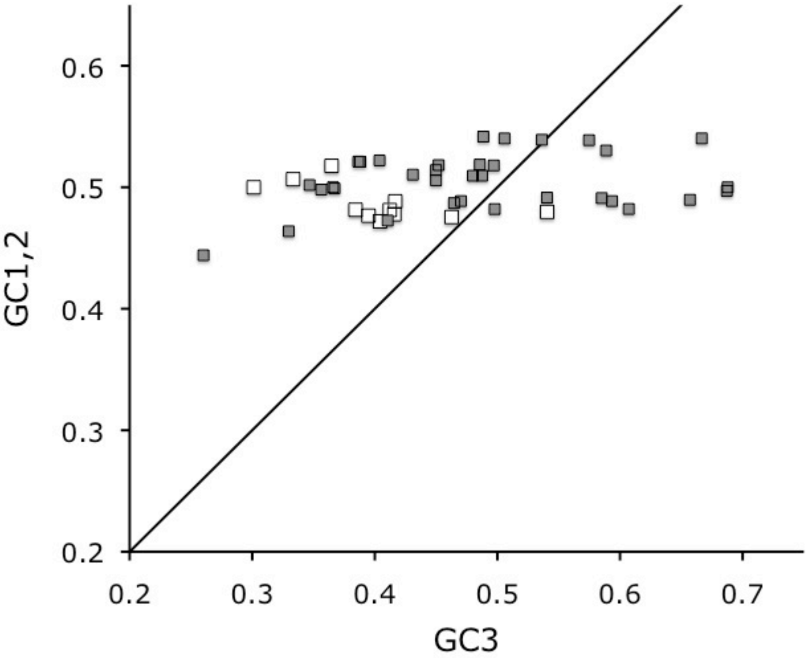


Figure 3. GC1,2/GC3 correlation for ssDNA (open squares) and dsDNA (grey squares) coliphage coat genes. Solid line indicates perfect correlation. Points above the line indicate genes deficient in GC3, points below denote genes enriched in GC3.

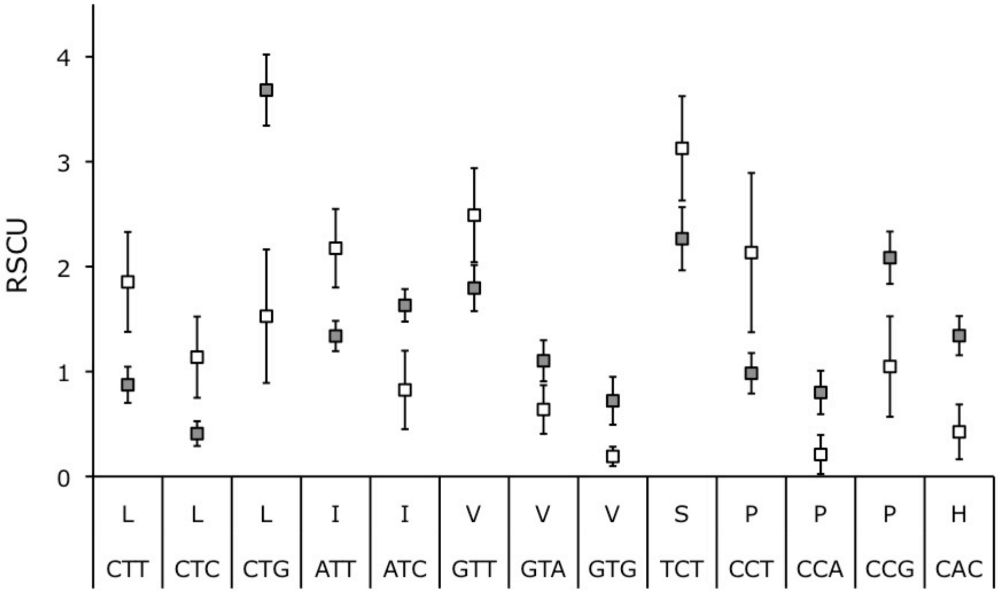


Figure 4. Mean RSCU values and 95% confidence intervals for individual codons with statistically significant differences in usage between ssDNA (open squares) and dsDNA (grey squares) coliphage coat or tail genes.

Calculation of RSCUs of coat genes in 28 ssDNA phages with a diverse host range confirmed this pattern: codons with thymine in the third position were extremely overrepresented (p<0.001) for six amino acids (A, D, G, I, T, V), and were significantly favored (p<0.012) in three more (H, P, S) (Figure 5). Only one of the remaining nine degenerate amino acids had a statistically preferred codon in ssDNA phages (GAA for E, p<0.01).

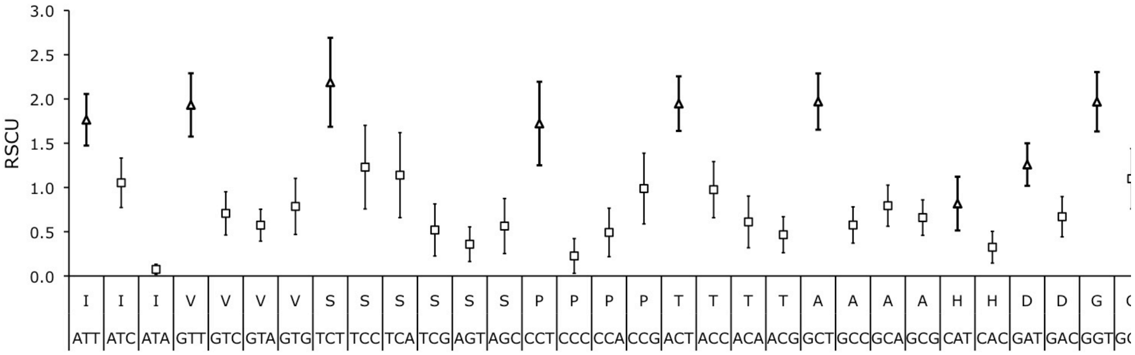


Figure 5. RSCU values and 95% confidence intervals for ssDNA phage coat gene codons that exhibited an NNT codon preference. Preferred NNT codons indicated by bold triangles, NNV codons indicated by squares.

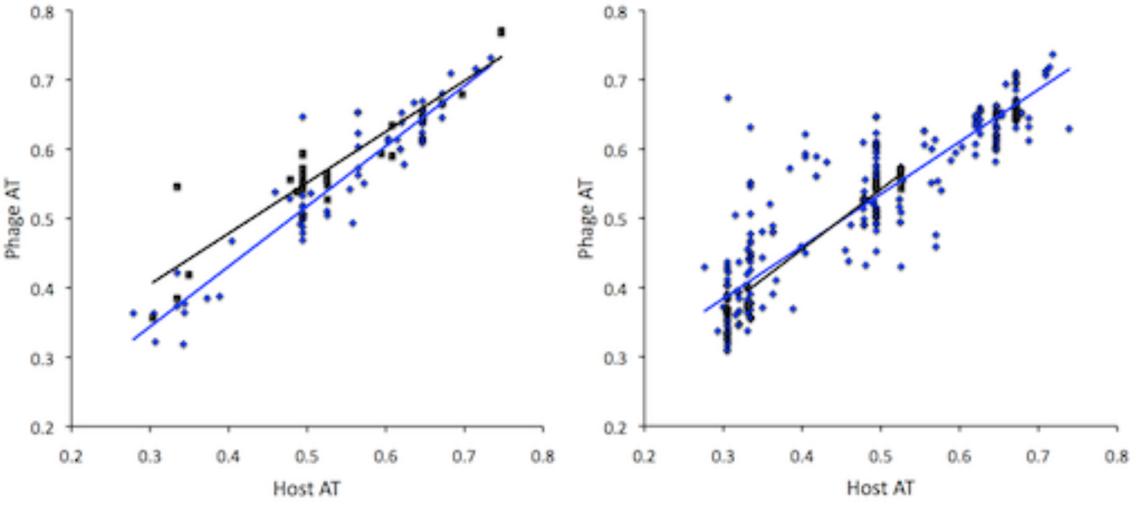
We subdivided our data set to separately examine the two morphologically distinct families of ssDNA phages, the *Inoviridae* and the *Microviridae*. Because inoviruses are frequently vertically transmitted and can productively infect their hosts without causing lysis, they might be under increased selective pressure to match the genomes of their more permanently associated hosts. RSCU comparisons revealed no consistent patterns associated with phage lifestyle. No difference in RSCU was evident for eleven of the sixteen NNT codons in these groups (Supplementary Figure 2).

Cytosines are comparatively unstable and readily undergo spontaneous deamination to uracil, resulting in C to T transitions after unrepaired replication (Poole et al., 2001). This spontaneous deamination occurs 100 times more frequently in ssDNA than dsDNA, resulting in a higher mutation rate at cytosines (Frederico et al., 1990) than at other bases in ssDNA phage (Cuevas et al., 2009). ssDNA phage genomes appear to spend more time truly single-stranded, as they do not experience consistent intra-strand base pairing or regular secondary structure formation while encapsidated (Benevides et al., 1991; Incardona et al., 1987; Shen et al., 1979; Tsuboi et al., 2010; Welsh et al., 1998; Wen et al., 1999). This causes ssDNA phages to more frequently have unpaired bases than ssRNA genomes, which are constrained by extensive stem-loop formation both in the cytosol and when encapsidated (Thurner, 2004).

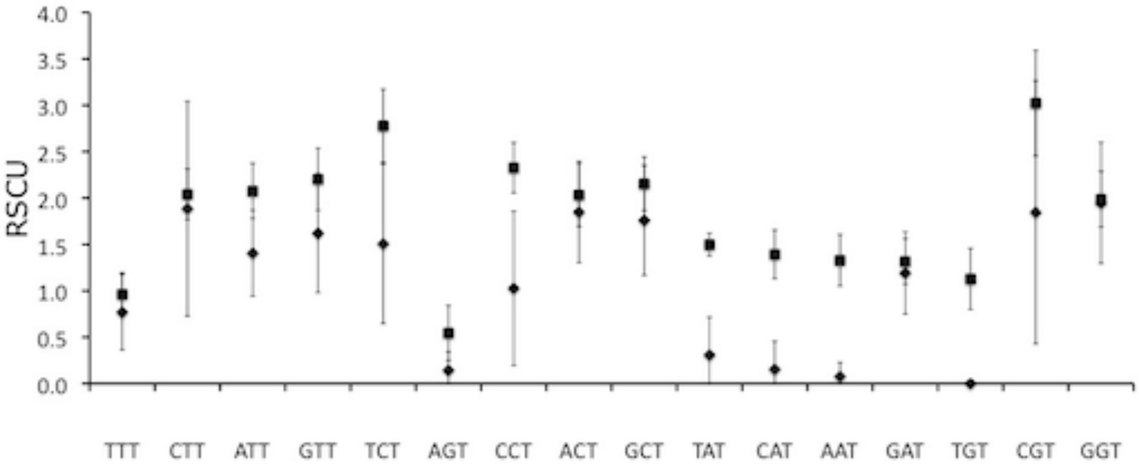
Any thymine-increasing bias does not appear to have a discernible effect on genomic nucleotide content relative to the phages' primary hosts. Rather, it is likely that cytosine transitions in the first or second positions are subject to strong purifying selection relative to the wobble position (Boyer et al., 1978; F.H.C, 1966; Holmes, 2003), and the signature of this mutational bias is only observed in the overabundance of thymine in the third position of synonymous ssDNA phage codons. The significant overrepresentation of NNT codons is strongly indicative of a biased mutational pressure acting in concert with strong selection against non-synonymous substitutions.

Genomic architecture (nucleic acid, segmentation, strandedness), while acknowledged as an important characteristic of virus taxonomy, is not typically included in broad-scale analyses of viral evolution. Instead, most comparisons focus within a single kind of virus (Jenkins and Holmes, 2003), and while many of these studies have provided insight into the codon usage biases of individual viruses, this is the first observation of a specific bias with a possible mechanistic explanation. Examining across two architectures, we saw strandedness play a critical role in the composition of phage genomes, and in determining the limits of ssDNA viral adaptation to their hosts.

Supplementary Figures



Supplementary Figure 1. Comparison of genomic AT content between ssDNA (black squares, black line) and dsDNA (blue diamonds, blue line) phage genomes and those of their hosts for phages previously analyzed by Xia and Yuen (2005) (left), and for phages added to GenBank since that analysis (right).



Supplementary Figure 2. Mean RSCU values and 95% confidence intervals for NNT codons in Microviridae (squares) and Inoviridae (diamonds).

Acknowledgements

We thank Dr. Eric Ho (Rutgers University) for use of his Python scripts. This work was supported by funds from the Rutgers School of Environmental and Biological Sciences, the New Jersey Agricultural Experiment Station and NSF MCB 1034927.

References

Adams, M.J., Antoniw, J.F., 2004. Codon usage bias amongst plant viruses. Arch. Virol. 149, 113-135.

Andersson, S.G., Sharp, P.M., 1996. Codon usage and base composition in Rickettsia prowazekii. J. Mol. Evol. 42, 525-536.

Antezana, M.A., Kreitman, M., 1999. The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences. J. Mol. Evol. 49, 36-43.

Aota, S., Ikemura, T., 1986. Diversity in G+C content at the third position of codons in vertebrate genes and its cause. Nucleic Acids Res. 14, 6345.

Bahir, I., Fromer, M., Prat, Y., Linial, M., 2009. Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. Mol. Syst. Biol. 5, 311.

Benevides, J.M., Stow, P.L., Ilag, L.L., Incardona, N.L., Thomas, G.J., 1991. Differences in secondary structure between packaged and unpackaged single-stranded DNA of bacteriophage phi X174 determined by Raman spectroscopy: a model for phi X174 DNA packaging. Biochemistry (Mosc.) 30, 4855-4863.

Bennetzen, J.L., Hall, B.D., 1982. Codon selection in yeast. J. Biol. Chem. 257, 3026-3031.

Bernardi, G., Bernardi, G., 1986. Compositional constraints and genome evolution. J. Mol. Evol. 24, 1-11.

Boyer, S.H., Scott, A.F., Kunkel, L.M., Smith, K.D., 1978. The proportion of all point mutations which are unacceptable: and estimate based on hemoglobin amino acid and nucleotide sequences. Can. J. Genet. Cytol. 20, 111-137.

Carbone, A., 2003. Codon adaptation index as a measure of dominating codon bias. Bioinformatics 19, 2005-2015.

Carbone, A., 2008. Codon bias is a major factor explaining phage evolution in translationally biased hosts. J. Mol. Evol. 66, 210-223.

Cuevas, J.M., Duffy, S., Sanjuán, R., 2009. Point mutation rate of bacteriophage PhiX174. Genetics 183, 747-749.

Curran, J.F., Yarus, M., 1989. Rates of aa-tRNA selection at 29 sense codons in vivo. J Mol Biol 209, 65-77.

Drake, J.W., 1991. A constant rate of spontaneous mutation in DNA-based microbes. Proc. Natl. Acad. Sci. U. S. A. 88, 7160-7164.

Duffy, S., Shackelton, L.A., Holmes, E.C., 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. 9, 267-276.

Duret, L., Mouchiroud, D., 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc. Natl. Acad. Sci. U. S. A. 96, 4482-4487.

F.H.C, C., 1966. Codon,Äîanticodon pairing: The wobble hypothesis. J. Mol. Biol. 19, 548-555.

Frederico, L.a., Kunkel, T.a., Shaw, B.R., 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry (Mosc.) 29, 2532-2537.

Hershberg, R., Petrov, D.A., 2008. Selection on codon bias. Annu. Rev. Genet. 42, 287-299.

Hiraoka, Y., Kawamata, K., Haraguchi, T., Chikashige, Y., 2009. Codon usage bias is correlated with gene expression levels in the fission yeast Schizosaccharomyces pombe. Genes Cells 14, 499-509.

Holmes, E.C., 2003. Patterns of Intra- and Interhost Nonsynonymous Variation Reveal Strong Purifying Selection in Dengue Virus. J. Virol. 77, 11296-11298.

Hyman, P., Abedon, S.T., 2010. Bacteriophage Host Range and Bacterial Resistance, in: Allen, I.L., Sima, S., Geoffrey, M.G. (Eds.), Advances in Applied Microbiology. Academic Press, pp. 217-248.

Ikemura, T., 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal

for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol 151, 389-409.

Ikemura, T., 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol 158, 573-597.

Incardona, N.L., Prescott, B., Sargent, D., Lamba, O.P., Thomas, G.J., 1987. Phage phi X174 probed by laser Raman spectroscopy: evidence for capsid-imposed constraint on DNA secondary structure. Biochemistry (Mosc.) 26, 1532-1538.

Jenkins, G.M., Holmes, E.C., 2003. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res. 92, 1-7.

Karlin, S., Mrázek, J., 1996. What drives codon choices in human genes? J. Mol. Biol. 262, 459-472.

Karlin, S., Mrázek, J., Campbell, a.M., 1998. Codon usages in different gene classes of the Escherichia coli genome. Mol. Microbiol. 29, 1341-1355.

Kawasaki, T., Nagata, S., Fujiwara, A., Satsuma, H., Fujie, M., Usami, S., Yamada, T., 2007. Genomic Characterization of the Filamentous Integrative Bacteriophages RSS1 and RSM1, Which Infect Ralstonia solanacearum. J. Bacteriol. 189, 5792-5802.

Lucks, J.B., Nelson, D.R., Kudla, G.R., Plotkin, J.B., 2008. Genome landscapes and bacteriophage codon usage. PLoS Comput. Biol. 4, e1000001.

Miller, E.S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T., Ruger, W., 2003. Bacteriophage T4 genome. Microbiol. Mol. Biol. Rev. 67, 86.

Ohama, T., Muto, a., Osawa, S., 1990. Role of GC-biased mutation pressure on synonymous codon choice in Micrococcus luteus, a bacterium with a high genomic GC-content. Nucleic Acids Res. 18, 1565-1569.

Ohkubo, S., Muto, a., Kawauchi, Y., Yamao, F., Osawa, S., 1987. The ribosomal protein gene cluster of Mycoplasma capricolum. Molecular & general genetics : MGG 210, 314-322.

Pedersen, S., 1984. Escherichia coli ribosomes translate in vivo with variable rate. The EMBO Journal 3, 2895-2898.

Poole, a., Penny, D., Sjöberg, B.M., 2001. Confounded cytosine! Tinkering and the evolution of DNA. Nat. Rev. Mol. Cell Biol. 2, 147-151.

Puigbò, P., Bravo, I.G., Garcia-Vallve, S., 2008. CAIcal: a combined set of tools to assess codon usage adaptation. Biol. Direct 3, 38.

Robinson, M., Lilley, R., Little, S., Emtage, J., Yarranton, G., Stephens, P., Millican, A., Eaton, M., Humphreys, G., 1984. codon usage can affect efficiency of translation of genes in Escherichia coli. Nucleic Acids Res. 12, 6663.

Sanjuán, R., Nebot, M.R., Chirico, N., Mansky, L.M., Belshaw, R., 2010. Viral mutation rates. J. Virol. 84, 9733-9748.

Serwer, P., Hayes, S.J., Zaman, S., Lieman, K., Rolando, M., Hardies, S.C., 2004. Improved isolation of undersampled bacteriophages: finding of distant terminase genes. Virology 329, 412-424.

Sharp, P.M., 1986. Molecular evolution of bacteriophages: evidence of selection against the recognition sites of host restriction enzymes. Mol. Biol. Evol. 3, 75-83.

Sharp, P.M., Li, W.-H., 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24, 28-38.

Sharp, P.M., Li, W.-H., 1987. The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15, 1281.

Sharp, P.M., Rogers, M.S., Mcconnell, D.J., 1985. Selection Pressures on Codon Usage in the Complete Genome of Bacteriophage T7. J. Mol. Evol., 150-160.

Shen, C.K., Ikoku, a., Hearst, J.E., 1979. A specific DNA orientation in the filamentous bacteriophage fd as probed by psoralen crosslinking and electron microscopy. J. Mol. Biol. 127, 163-175.

Shields, D.C., Sharp, P.M., Higgins, D.G., Wright, F., 1988. "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. 5, 704-716.

Sueoka, N., Kawanishi, Y., 2000. DNA G+C content of the third codon position and codon usage biases of human genes. Gene 261, 53-62.

Thomas, J.a., Rolando, M.R., Carroll, C.a., Shen, P.S., Belnap, D.M., Weintraub, S.T., Serwer, P., Hardies, S.C., 2008. Characterization of Pseudomonas chlororaphis myovirus 201varphi2-1 via genomic sequencing, mass spectrometry, and electron microscopy. Virology 376, 330-338.

Thurner, C., 2004. Conserved RNA secondary structures in Flaviviridae genomes. J. Gen. Virol. 85, 1113-1124.

Tsuboi, M., Tsunoda, M., Overman, S.A., Benevides, J.M., Thomas, G.J., 2010. A Structural Model for the Single-Stranded DNA Genome of Filamentous Bacteriophage Pf1. Biochemistry (Mosc.) 49, 1737-1743.

Welsh, L.C., Marvin, D.A., Perham, R.N., 1998. Analysis of X-ray diffraction from fibres of Pf1 Inovirus (filamentous bacteriophage) shows that the DNA in the virion is not highly ordered. J. Mol. Biol. 284, 1265-1271.

Wen, Z.Q., Armstrong, A., Thomas Jr, G.J., 1999. Demonstration by ultraviolet resonance Raman spectroscopy of differences in DNA organization and interactions in filamentous viruses Pf1 and fd. Biochemistry (Mosc.) 38, 3148-3156.

Wong, E.H.M., Smith, D.K., Rabadan, R., Peiris, M., Poon, L.L.M., 2010. Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus. BMC Evol. Biol. 10, 253.

Xia, X., 2007. An Improved Implementation of Codon Adaptation Index. Evolutionary Bioinformatics, 53-58.

Xia, X., Yuen, K.Y., 2005. Differential selection and mutation between dsDNA and ssDNA phages shape the evolution of their genomic AT percentage. BMC genetics 6, 20.

Chapter 2

Base Composition and Translational Selection are Insufficient to Explain Codon Usage Bias in Plant Viruses

Abstract

Viral codon usage bias may be the product of a number of synergistic or antagonistic factors, including genomic nucleotide composition, translational selection, genomic architecture, and mutational or repair biases. Most studies of viral codon bias evaluate only the relative importance of genomic base composition and translational selection, ignoring other possible factors. We analyzed the codon preferences of ssRNA (luteoviruses and potyviruses) and ssDNA (geminiviruses) plant viruses that infect translationally distinct monocot and dicot hosts. We found that neither genomic base composition nor translational selection satisfactorily explains their codon usage biases. Furthermore, we observed a strong relationship between the codon preferences of viruses in the same family or genus, regardless of host or genomic nucleotide content. Our results suggest that analyzing codon bias as either due to base composition or translational selection is a false dichotomy that obscures the role of other factors. Constraints such as genomic architecture and secondary structure can and do influence codon usage in plant viruses, and likely in viruses of other hosts.

Introduction

All organisms exhibit some degree of codon usage bias (CUB), the unequal usage of synonymous codons (Aota and Ikemura, 1986; Shields et al., 1988). Codon bias may vary among genes of the same organism, which is associated with factors like asymmetrical mutation pressures or tissue-specific gene expression, but is relatively uniform within the most highly expressed genes (Bailly-Bechet et al., 2006; Bennetzen and Hall, 1982; Camiolo et al., 2012; Hershberg and Petrov, 2008; Hiraoka et al., 2009; Karlin et al., 1998; Lobry and Sueoka, 2002; Sharp and Li, 1986). CUB is often explained as the product of two potentially competing factors: genomic base composition and translational selection (Table 1). In the absence of other mutational and selective pressures, CUB should result from the genomic frequency of A, C, G and T being reflected in third positions. When CUB diverges from the null hypothesis of genomic nucleotide content, translational selection—selection for optimal speed and accuracy of translation—is routinely invoked. Translational selection should exert an influence on CUB because preferred codons tend to correlate with the most common tRNAs (Ikemura, 1981; Ikemura, 1982), allowing for faster, yet accurate, codon recognition and translation of highly expressed genes (Curran and Yarus, 1989; Pedersen, 1984). However, genomic composition and translational selection need not be acting antagonistically on CUB, and sequences can show CUB distinct from that predicted by either force.

Table 1. Possible explanations for codon usage bias when the codon usage bias (CUB) of a gene of interest match or fail to match the genomic base composition and relative synonymous codon usage (RSCU) of a set of reference genes. For viruses, the comparison would be between the CUB of the viral genes and the CUB of their hosts.

		Base Composition	
		Match	No Match
RSCU	Match	Both?	Translational Selection
	No Match	Base Composition	Undetermined

When genomic base composition and known preferred codons correlate with observed CUB, both are potentially influencing CUB, and we cannot distinguish the relative strength of the forces. When observed CUB conflicts with the known preferred codons but adheres to genomic nucleotide content, the null hypothesis of overall base composition cannot be rejected, but translational selection can. Conversely, when preferred codons and observed CUB align, but CUB differs from that predicted by genomic base content, the null hypothesis can be rejected and translational selection deemed a more likely explanation. In the fourth case, neither overall base composition nor translational selection appears to be driving the observed CUB. This could be the result of direct conflict between the two forces yielding an intermediate state (i.e., the genome is enriched in adenine and suppresses cytosine, but the preferred codons tend to end in cytosine at the expense of NNA codons). Alternatively, another factor or factors must be influencing CUB. However, the vast majority of studies into the causes of CUB frame the question in terms of genomic base composition vs. translational selection, which precludes the consideration of additional important factors (Cai et al., 2009; Jia et al., 2008). Considerations such as species-specific nmer promotion and suppression (*e.g.*, GATC for methyl-directed mismatch repair in *E. coli* (Au et al., 1992), CpG in mammalian genomes (Krieg et al., 1995)) are known to affect CUB, but are rarely considered in analyses of codon usage.

This framework can be applied to studies of viral codon bias. Viruses with well characterized hosts are ideal systems in which to explore the forces shaping CUB because their genomic biases can be calculated from their viral genomes, but the hosts' CUB reveal the translationally preferred codons. Viruses should experience translational selection to match the CUB of their hosts, as this should allow for faster translation of highly expressed viral genes, and consequently more rapid viral replication. It was recently documented that viruses with highly deoptimized CUB suffer a fitness cost (Bull et al., 2012). However, surveys examining viral CUB have indicated that not all viruses are equally able to match their hosts' codon preferences, and that this may be correlated with viral genomic architecture (Jenkins and Holmes, 2003). For instance, we previously demonstrated that double-stranded (dsDNA) coliphages were significantly better matched to *Escherichia coli*'s CUB than single-stranded (ssDNA) coliphages, because ssDNA phages had a preference for NNT codons, regardless of the hosts' preferred codon usage (Cardinale and Duffy, 2011).

To further investigate the evolutionary forces shaping viral codon usage, we investigated patterns of CUB in plant viruses with distinct genomic architectures. Plant viruses offer a unique opportunity to examine CUB because plant virus families often include members that only infect monocot hosts, and others that only infect eudicot hosts—two distinct translational environments. Monocots tend to have GC biased genes (53–56%), while eudicot genes generally have lower GC content (40-45%) (Wang and Roossinck, 2006). Monocots exclusively prefer G- and C-ending codons, while eudicots prefer a combination of G- and T-ending codons in their most highly expressed genes (Table 2). These divergent hosts allow the strength of translational selection pressures to be compared among related viruses.

Table 2. Preferred codons in monocots and eudicots. Preferred codons are those with relative synonymous codon usage (RSCU) values that significantly exceed those of all synonymous codons (p<0.05, Bonferroni-corrected 2-tailed t-tests).

	Monocots			Eudicots		
NNC	tac	aac	ccc	tac	aac	
	ctc	acc	gac			
	atc	gcc	tgc			
	tcc	tac	cgc			
	agc	cac	ggc			
NNG	ttg	aag	cag	ttg	aag	cag
	gag	agg		gag		
NNT				ctt	tct	gct
				gtt	gat	cgt

We analyzed three large groups of arthropod-vectored plant viruses: the positive sense ssRNA genus *Potyvirus* and family *Luteoviridae,* and the ssDNA family *Geminiviridae*. *Potyvirus* and *Geminiviridae* contain a comparable number of species with at least 15 sequences available for analysis (22 and 24, respectively). There were fewer appropriate *Luteoviridae* for analysis (8), but similar to the *Geminiviridae*, monocot- and dicot- infecting luteoviruses are organized into separate genera. These three groups differ in their genomic architectures: the filamentous potyviruses have a linear ~10kb genome that is expressed as a polyprotein, luteoviruses contain a linear genome of 5.3- 5.7kb that is translated from subgenomic RNAs, and geminiviruses have one or two circular, ~2.7kb, ambisense genomic segments that are transcribed by host enzymes (King et al., 2011).

Unlike cellular organisms, which share related genes across extremely divergent clades, which can be used as the basis for phylogenies (Woese et al., 1990), very few functionally analogous viral genes are found in divergent taxa. We chose to examine the coat/capsid protein (CP) gene, a large ORF that is shared (though not homologous) among the three viral groups. While the CPs in some plant viruses serve the dual role of capsid and movement proteins (Rojas et al., 1997), these factors only constrain amino acid usage, and should not impact synonymous codon usage. Similarly, the CPs of vectored viruses are under more strict selection against amino acid substitutions than those of non-vectored viruses (Chare, 2004), but as these arthropod-borne viruses are not expressing genes in the vector, it should not affect their codon bias. Therefore, analysis of CP genes best facilitates comparisons of results between the ssDNA and ssRNA viruses in this study.

Additional analyses were required in *Geminiviridae*. The potyvirus CP and the luteovirus CP are each considered monophyletic; the monocot- and dicot-infecting viral sequences within each group once shared a common ancestral sequence. The monophyly of the CP in the *Geminiviridae* is assumed based on its unusual capsid shape (Krupovic et al., 2009), but the protein sequence of the ORF is highly divergent between begomoviruses and mastreviruses. Consequently, we also analyzed the CUB of the replication-associated gene (Rep), which is encoded in the complementary sense, for the geminiviruses. There is strong phylogenetic evidence for their Reps to be descended from a common ancestor (Martin et al., 2011; Rosario et al., 2012). While their CPs may be useful for comparisons to the RNA viruses, comparisons of the Rep CUB within *Geminiviridae* may be more appropriate, and comparable to analyses of the homologous CPs within each RNA virus family (Krupovic et al., 2009; Varsani et al., 2009).

We compared the relative synonymous codon usage (RSCU (Sharp and Li, 1986)) of monocot- and eudicot-infecting members of each group to their hosts, to each other, and to viral genomic nucleotide composition to assess the relative importance of host codon preferences in viral CUB. Our results were surprisingly variable for viruses infecting common hosts, and demonstrate that pressures beyond base composition and translational selection affect CUB in ssDNA and ssRNA plant viruses.

Methods

Host codon usage bias

Codon preferences in highly expressed genes for five monocot and six eudicot plants were determined using the RSCU data from Wang and Roossinck (2006). Average monocot and eudicot RSCU was calculated for each codon, and preferred codons were defined by Bonferroni-corrected two- tailed t-tests (Microsoft Excel) of average RSCU for synonymous codons. Each set of redundant codons was analyzed individually; no comparisons were made between non-redundant codons. For these analyses, codons for the six-fold degenerate amino acids (L, R, S) were divided into two-fold and four-fold redundant groups, for which RSCU values were calculated independently. This was done because the two groups of codons for these amino acids differ at non-synonymous sites and are consequently recognized by different groups of tRNA species, making it inappropriate to treat them as a single set of redundant codons.

Plant virus datasets

All available complete CP sequences of luteoviruses and potyviruses were collected from GenBank between March and May of 2012. Only species with 15 or more full CP gene sequences were analyzed, and sequences with ambiguous nucleotides were excluded. Complete CP and complete Rep gene sequences of Geminiviruses (monocot-infecting mastreviruses and eudicot-infecting begomoviruses) were downloaded from GenBank between January and April of 2012. As with the ssRNA viruses, only species with at least 15 full gene sequences were analyzed, and sequences with ambiguous nucleotides were excluded. Consequently, some geminivirus species could be included in one of our analyses (CP analysis) but not the other (Rep analysis). In total, we analyzed 1285 geminivirus Rep gene sequences, 1481 geminivirus, 1210 potyvirus, and 315 luteovirus CP gene sequences.

Base composition as a null hypothesis

All sequences were formatted for analysis using ReadSeq (http://www-bimas.cit.nih.gov/molbio/readseq). CAICal (Puigbò et al., 2008) was used to calculate the viral base composition. Reference sequences for each viral species were collected from GenBank on June 12, 2012. Sequences were formatted with ReadSeq, and CAICal was used to determine overall and site-specific base composition for each sequence. Observed third position nucleotide counts were averaged for each species. Expected third position nucleotide counts were computed for each gene/species combination we analyzed based on the genomic nucleotide frequencies of the species' reference genome and the length of the ORF in the reference genome. Chi-square tests were used to evaluate the differences between these observed and expected counts, with three degrees of freedom (MS Excel). In total, sixty-seven chi-square tests were carried out: one on the CP gene of each potyvirus, luteovirus, and geminivirus we examined, and one on the Rep gene of each geminivirus we analyzed.

Plant virus codon usage biases

Viral RSCU calculations were as for the plant hosts. RSCU values were calculated for each sequence in each viral genus (in the case of the potyviruses, monocot-infecting and eudicot-infecting). Then, mean RSCU values were calculated for each viral genus/group. Preferred codons were again defined by Bonferroni-corrected two-tailed t-tests of average RSCU for synonymous codons, and determined separately for monocot- and dicot-infecting members of each viral group.

Comparison with host CUB

Average RSCU values for the monocot- and eudicot-infecting viruses of each viral group were compared to those of their respective hosts to determine the correlation between host and viral CUB. Average RSCU of monocot-infecting and dicot-infecting viruses within each group were also compared to each other to determine if the correlation among related viruses was stronger than the correlation to their respective hosts. RSCU between two groups was classified as uncorrelated ($r<0.50$), moderately correlated ($0.50\leq r<0.70$), or strongly correlated ($r\geq 0.70$). Translational selection was rejected when host and virus RSCU were uncorrelated, but considered in cases of moderate or strong correlation.

Results

Monocots and Eudicots exhibit divergent CUB

Plant codon preferences varied considerably between monocots and eudicots. The monocots analyzed exclusively preferred C- and G-ending codons in their highly expressed genes; fifteen of their overrepresented codons were C-ending, while the remaining five were G-ending (Table 2). These patterns were consistent with the codon preferences of all monocot genes (Wang and Roossinck, 2006). Conversely, eudicots preferred a combination of NNT (six) and NNG (four) codons in their most highly expressed genes, in addition to two NNC codons (Table 2), which also agreed with their overall codon preferences. In all cases where plants preferred an NNG codon, no pyrimidines were possible in the third position (they were two-fold redundant amino acids, or the two-fold redundant portion of six-fold redundant amino acid codons). The RSCU of the highly expressed genes in the monocots and eudicots (Wang and Roossinck, 2006) were strongly correlated for A or T-ending codons (r=0.93), and for G or C-ending codons (r=0.82, Figure 1). However, best-fit lines for the two groups of codons both differ from a line with a slope of 1 through the origin, indicative of divergent codon preferences. Consequently, monocots and eudicots represent distinct translational environments, and should exert dissimilar translational selection pressures on the CUB of their respective viruses.

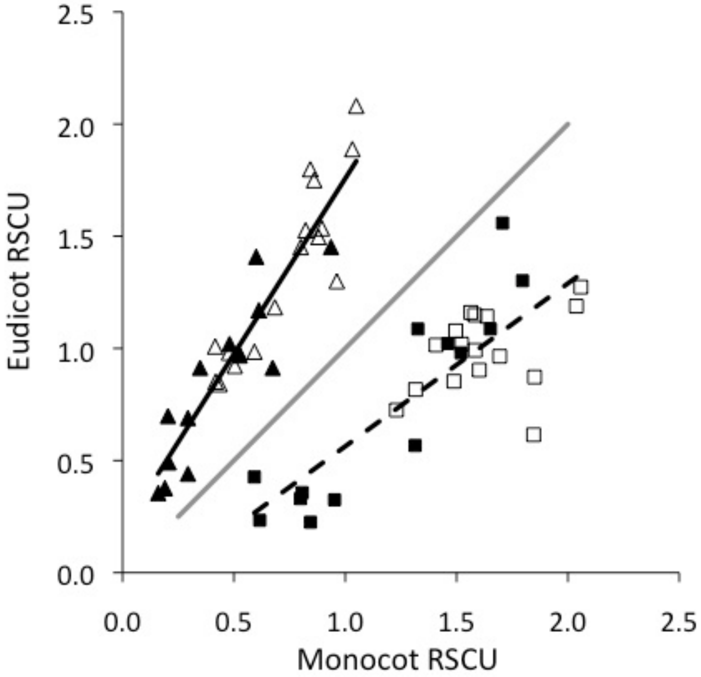


Figure 1. Relative synonymous codon usage (RSCU) correlation between monocot and eudicot highly expressed genes. Triangles represent A/T, squares are C/G, open symbols are pyrimidines, closed are purines. Solid line is best fit for A/T-ending codons (r=0.93), dashed is for G/C-ending (r=0.82). The grey line has a slope of 1 through the origin.

Base composition does not explain most CUB in plant viruses

The potyviruses showed a consistent pattern of elevated adenine in their genomes, regardless of host, and also contained correspondingly lower levels of cytosine and guanine. The third position nucleotide content in CP genes differed significantly from that of the overall genome in every potyvirus we examined (chi-square tests, p<0.05). Luteoviruses showed consistent genomic base composition, having slightly elevated genomic adenine content, and relatively equitable use of cytosine, guanine, and thymine. Third position base frequencies were also consistent regardless of host, but differed significantly from genomic nucleotide composition in most luteoviruses. Third position base usage in two of the four eudicot-infecting luteoviruses did not differ significantly from genomic base content (chi-square tests, p>0.1). In the two remaining eudicot-infecting, and all four monocot-infecting luteoviruses, third positions diverged significantly from the genomes (chi-square tests, p<0.05). These findings indicate that genomic base composition is a poor predictor of CUB in luteoviruses and potyviruses.

Average third position base composition of CP genes in begomoviruses (eudicot-infecting geminiviruses) and mastreviruses (monocot-infecting geminiviruses) also varied greatly from their respective genomic nucleotide contents (18 begomoviruses and 4 mastreviruses, chi-square tests, p<0.001). Base composition of synonymous sites in the Rep genes of all begomoviruses (n=14) and two out of three mastreviruses also diverged substantially from overall genomic nucleotide content (chi-square tests, p<0.05). As is the case for the ssRNA viruses, these results strongly suggest that genomic base composition does not drive CUB in geminiviruses.

RNA virus CUB is independent of host use

All potyviruses had somewhat similar codon preferences, independent of host: monocot- and eudicot-infecting potyviruses both generally preferred A- and T-ending codons (Table 3). Luteoviruses, both monocot- and eudicot-infecting, exhibited preferences for fewer codons overall, but tended to favor NNC codons. Despite this overall similarity they shared only two preferred codons (Table 3).

Table 3. Preferred codons in luteoviruses and potyviruses infecting monocot and eudicot hosts. Preferred codons are those with relative synonymous codon usage (RSCU) values that significantly exceed those of all other synonymous codons (p<0.05, Bonferroni-corrected 2-tailed t-tests).

	Potyviruses		Luteoviruses	
Host	monocot	eudicot	monocot	eudicot
NNA	tca cca gca	tca cca gca		aga
	aga aaa aca	aga gaa gga		
		caa		
NNC	tgc cac		tgc ttc gac	tgc ttc gtc
			tac	atc ctc
NNT	tat aat gat	tat aat gat		
	ttt ctt gtt	ttt		
NNG		ttg	agg	

Eudicot-infecting potyvirus RSCU was moderately correlated with eudicot RSCU (r=0.55), and monocot-infecting RSCU was actually weakly anti-correlated with monocot RSCU (r=-0.29, Figure 2a). Surprisingly, monocot-infecting potyvirus RSCU correlated with eudicot RSCU nearly as well as eudicot-infecting potyviruses (r=0.46). Monocot- and eudicot-infecting luteovirus RSCU weakly correlated with that of their hosts (r=0.29 and 0.16, respectively, Figure 2b).

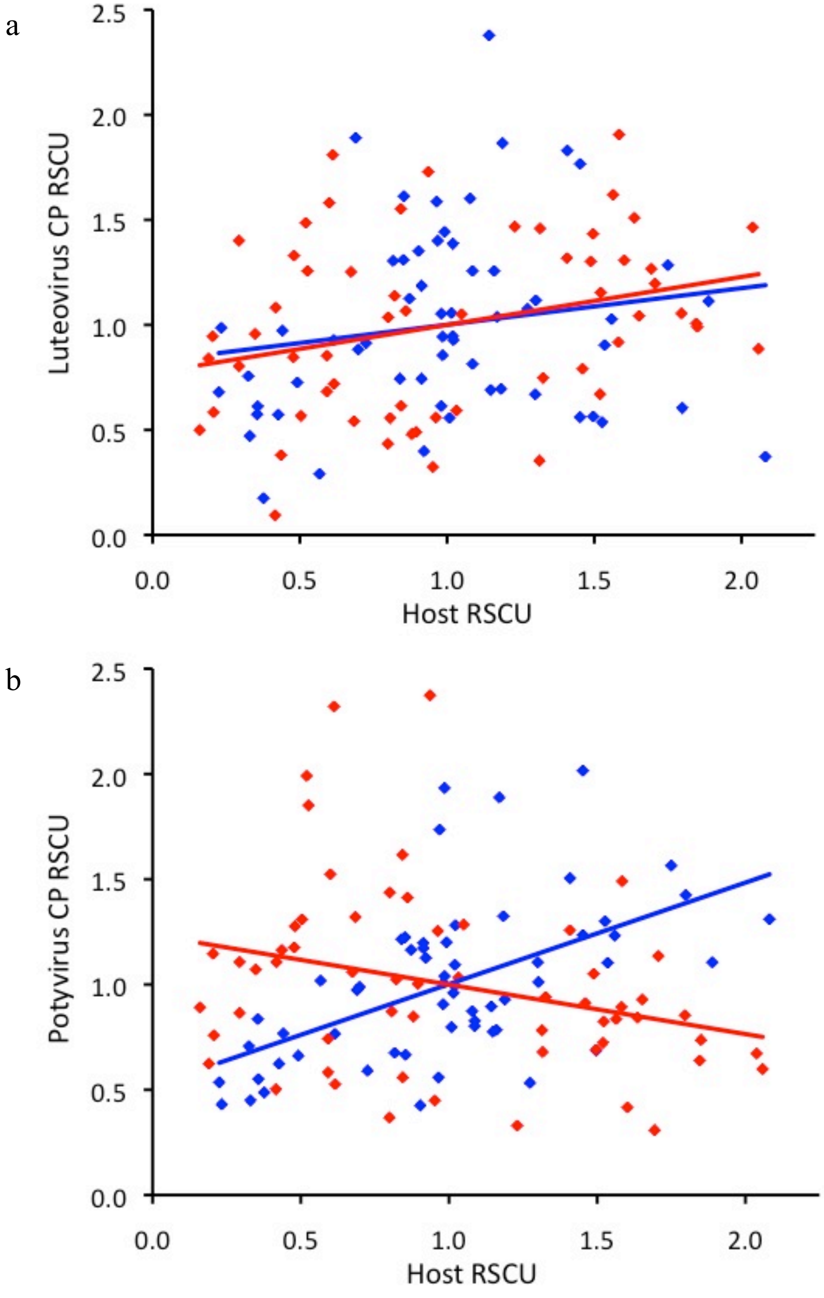


Figure 2. Correlation between host and virus coat/capsid protein (CP) relative synonymous codon usage (RSCU) for monocot-infecting (red) and eudicot-infecting (blue) **(a)** potyviruses and **(b)** luteoviruses.

ssDNA virus RSCU does not indicate strong translational selection

ssDNA dicot-infecting begomovirus CP genes exhibited a strong preference for NNT codons, while begomovirus Rep sequences strongly favored

NNA codons (Table 4). In the monocot-infecting mastreviruses, CP genes preferentially used C- and G-ending codons, but the Rep sequences did not exhibit a specific preference; overrepresented codons ended in all four bases (Table 4). Begomovirus genomes are ambisense; genes are encoded in the coding and complimentary sense (Gutierrez, 1999). The coding sequence of the Rep gene is complimented on the virion strand. As a consequence, third positions in this gene are present as the first base of anti-codons in the single-stranded viral genome. Therefore, these findings indicate begomovirus genomes are enriched for thymine at synonymous sites in both the CP ORF (with T-ending codons) and Rep ORF (with T-beginning anticodons).

Table 4. Preferred codons in mastrevirus and begomovirus Rep and CP genes. Preferred codons are those with relative synonymous codon usage (RSCU) values that significantly exceed those of all other synonymous codons (p<0.05, Bonferroni-corrected 2-tailed t-tests).

	Mastreviruses		Begomoviruses	
ORF	Rep	CP	Rep	CP
NNA	aaa		aaa caa gga aca cca gaa aga	
NNC	tac	ttc gcc gac	ttc ctc tgc	ttc ccc
NNT	cat cgt	agt	cat aat gat	cat cgt aat gat act tgt att ggt gtt tat
NNG	ttg agg	ttg agg gag aag cag ctg	ttg	ttg agg gag aag

In both begomoviruses and mastreviruses, the RSCU of their CP genes was moderately correlated to the RSCU of the highly expressed genes of their respective hosts (r=0.69 for begomoviruses, 0.53 for mastreviruses, Figure 3a). However, the Rep gene was not well correlated to their host RSCU in either begomoviruses (r=0.38) or mastreviruses (r=0.06, Figure 3b). Mastrevirus Rep RSCU actually matched that of eudicots better than the begomovirus Rep RSCU (r=0.71).

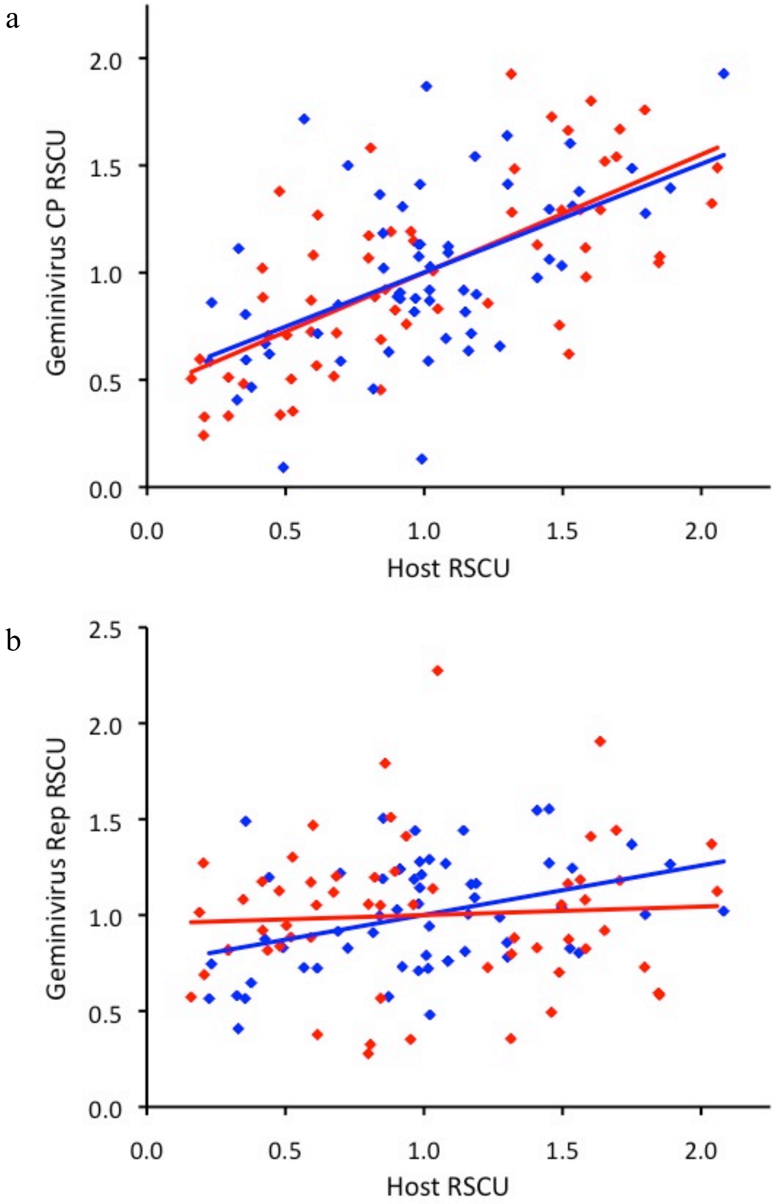


Figure 3. Correlation between relative synonymous codon usage (RSCU) of **(a)** geminivirus CP and **(b)** geminivirus Rep and host RSCU for eudicot-infecting begomoviruses (blue) and monocot-infecting mastreviruses (red).

The conservation of codon usage between monocot- and eudicot-infecting viruses within each group varied significantly. Potyvirus RSCUs were strongly correlated to each other (r=0.90), despite their hosts having divergent preferences (Figure 4a). Similar to the potyviruses, the two luteovirus groups exhibited a moderate correlation to each other (r=0.66, Figure 4a). Begomovirus and mastrevirus RSCUs were uncorrelated in the CP ORF (r=0.16), but moderately correlated in the Rep ORF (r=0.51, Figure 4b).

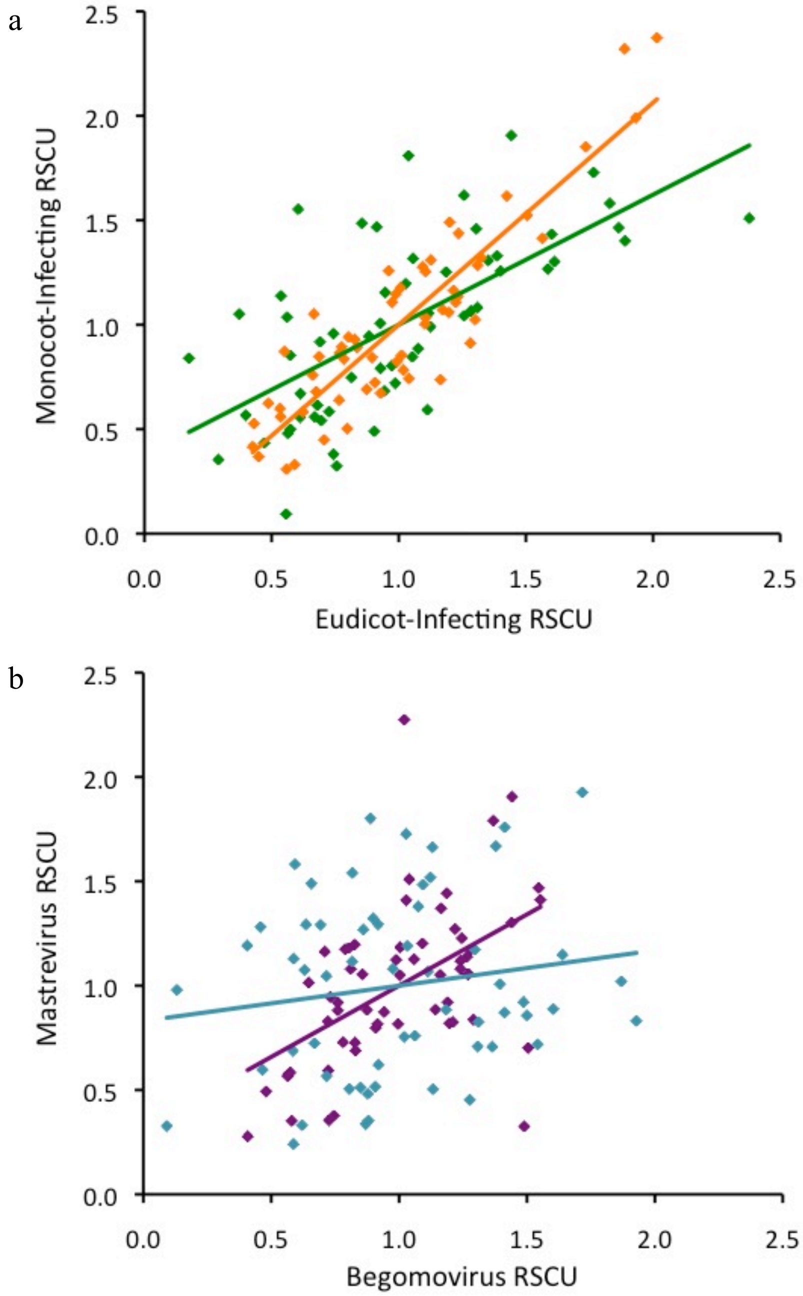


Figure 4. Correlation of relative synonymous codon usage (RSCU) between **(a)** monocot and eudicot-infecting potyvirus CPs (orange), luteovirus CPs (green) and **(b)** begomo- and mastrevirus CPs (aqua) and begomo- and mastrevirus Reps (purple).

Discussion

Neither base composition nor translational selection explains our results

The null hypothesis, that synonymous codon usage is purely a function of the nucleotide base frequencies in the viral genome, was insufficient to explain codon preferences in all of the groups we examined. The discord between genomic frequencies and third position frequencies did not often manifest as more equitable nucleotide use in the third position; instead, codon use was more biased than genomic nucleotide frequencies in several cases (eudicot-infecting potyviruses, begomovirus CP). Similarly, the alternative hypothesis of translational selection failed to explain much of the plant virus CUB. Dicot-infecting potyviruses and geminivirus CP CUB were moderately correlated with their host CUB, but we did not find as high a correlation as in phage (Cardinale and Duffy, 2011) or human RNA viruses (Jenkins and Holmes, 2003).

The most common methods of assessing the relationship between genomic base composition and CUB involve using GC3, the GC content of the third codon position, as a measure of codon bias. GC3 is then compared to GC1,2, the GC content of the first and second codon positions, or ENC, the effective number of codons (Adams and Antoniw, 2004; Fuglsang, 2005). Both measures vary along predictable lines or curves when base composition drives CUB. By these measures, CUB in many viruses is strongly affected by overall base composition (Adams and Antoniw, 2004; Jiang et al., 2008; Liu et al., 2010). However, metrics that group AT and GC are unable to account for the over- or underrepresentation of a specific nucleotide in the third position at the expense of its complement, which is of special concern when analyzing CUB in single-stranded viruses. In the begomoviruses we analyzed, the average GC3 is approximately equal in the CP, Rep, and whole genome, but the CP (and genome) favor guanine while cytosine is overrepresented in the Rep. Consequently, our methods require a higher degree of agreement between the overall genomic base content and gene third position for base composition to be considered a plausible explanation for CUB.

Furthermore, many studies of viral CUB do not explicitly evaluate other factors that can drive codon bias. Rather, the null hypothesis is tested—does CUB follow the predicted relationship between GC3 and GC1,2 or ENC—and if not rejected, the effects of translational selection and other possible factors are not subsequently analyzed (Cai et al., 2009; Jia et al., 2008; Jiang et al., 2008; Liu et al., 2010). Other studies attribute the rejection to translational selection, but fail to consider additional factors (Wang et al., 2011; Wang et al., 2010; Xu et al., 2008).

Possible alternative explanation for CUB in plant viruses

Genes of propagative arthropod-vectored viruses (those that replicate within their vectors) should be under dual selective pressures to maximize replication speed within their plant hosts and their vectors. Consequently, vector codon preferences could influence codon bias in these viruses. However, the potyviruses and luteoviruses are nonpropagative (Gray and Banerjee, 1999), and while the evidence is more ambiguous in geminiviruses, they are generally considered nonpropagative as well (Andret-Link and Fuchs, 2005; Power, 2000), so translational selection is not acting on these viral genomes in their respective vectors. There is evidence of heightened purifying selection on capsid structure in vectored RNA viruses due to specific interactions between CP and vector (Chare, 2004), but in nonpropagative viruses, this pressure is independent of translation kinetics, instead acting solely on amino acid sequence. Therefore, translational selection within arthropod vectors cannot explain the observed CUB.

As potyviruses have high mutation rates (Sanjuan et al., 2009), have diverged over at least thousands of years (Gibbs et al., 2008), and the different species analyzed were at least 25% divergent by nucleotide (King et al., 2011), it is impossible that the common synonymous codon usage we observed is an accident of recent fixation. It is similarly unlikely that a recent host-shift from eudicots (to which potyvirus CUB is better matched) brought recently diverged potyviruses into monocots (Gibbs et al., 2008). Luteoviruses (Pagan and Holmes, 2010) and geminiviruses (Duffy and Holmes, 2008; Duffy and Holmes, 2009; Harkins et al., 2009) evolve at similarly high speeds, so it is unlikely that these correlations are due to accidental historical contingency. It would further be expected that third positions would be saturated after thousands, if not millions of years of divergence (Lefeuvre et al., 2011). Given these factors, it is most likely that the correlation among monocot- and eudicot-infecting members of each group is due to a similar set of pressures affecting CUB of each group as a whole.

One possible factor that may influence codon bias in ssRNA viruses is selective constraints on secondary structure. ssRNA viral genomes often contain complex secondary structures that are important for replication or gene expression (Hofacker et al., 2004). Disruption of these structures can inhibit one or both processes, reducing viral fitness. Substitutions are often observed in pairs: an initial mutation and a compensatory mutation that restores base pairing across stems in stem-loop structures, for instance (Hofacker et al., 1998). These factors should manifest as more constrained codon usage at specific sites, though the effects on overall codon usage are ambiguous.

The begomovirus CP, which strongly preferred NNT codons, aligned well with the preference of their eudicot hosts for T-ending codons and correlated strongly with host RSCU. Despite the significant differences between third position and genomic base content, these results also indicate the potential importance of the thymine enrichment of high-AT begomovirus genomes. Therefore, it is tempting to explain these data as the result of the combination of compositional constraints and strong translational selection, even if third position base use significantly differed from that of the entire genome. Conversely, begomovirus Rep sequences have different preferences and demand a different explanation. CUB is not explained by base composition, but the prevalence of A-ending codons and the weak correlation (r=0.37, Figure 4b) between host and virus RSCU suggests weak translational selection. It is unlikely that these two genes are subject to such divergent pressures that they would exhibit such inverse biases, as CUB tends to be similar within species (Grantham et al., 1980). However, neither base composition nor translational selection is sufficient to explain begomovirus CUB.

Similarly, CUB in the mastreviruses is not easily explained. In particular, the Rep has no well- defined codon preferences, which is not predicted by the genomic nucleotide composition, translational selection for their hosts' CUB, nor an antagonistic relationship between the two. Consequently, these two factors alone are not sufficient to explain mastrevirus CUB.

The ssDNA architecture of geminiviruses provides a possible explanation for their CUB. ssDNA is prone to rapid cytosine deamination to uracil (Frederico et al., 1990), and this process may explain the preference for T-ending codons in ssDNA phages, even in hosts with low AT% (Cardinale and Duffy, 2011). If this process also affects eukaryotic ssDNA viruses, we would expect a very different CUB profile compared to that which is determined by only base content and translational selection. Specifically, strong translational selection predicts uniform codon usage in both CP and Rep, but a strong, biased mutational pressure predicts the begomovirus preference for A-ending codons in the Rep sequences, given that they are encoded in the negative sense. C→T transitions may be tolerated only at synonymous sites, resulting in an overabundance of thymine in the genomic sequence, and a corresponding preference for adenine in the Rep coding sequence. When viewed as they are encoded in the genome, begomovirus CP and Rep nucleotide preferences at synonymous sites are remarkably consistent: both strongly prefer T-ending codons/T-beginning anticodons, suggesting that this biased mutational pressure may contribute to geminivirus CUB. A recent study of the ssDNA porcine circovirus also shows a preference for T-ending codons that differs from the codon preferences of their swine hosts, and is not due to genomic composition (Liu et al., 2012).

We believe it is very likely that a biased C→T mutational pressure affects eukaryotic ssDNA viruses. Eudicot-infecting begomoviruses are known to exhibit a long-term C→T substitution bias (Duffy and Holmes, 2008; Duffy and Holmes, 2009). Additionally, ssDNA phages typically exhibit little secondary structure (Benevides et al., 1991; Incardona et al., 1987; Shen et al., 1979; Tsuboi et al., 2010; Welsh et al., 1998; Wen et al., 1999), and while very limited degrees have been documented in eukaryotic ssDNA viruses (Sun et al., 2009), it has not been observed in the ORFs we examined here. Consequently, these genomes are unconstrained by structural constraints at synonymous sites, and, because unpaired DNA is 100 times more susceptible to oxidative cytosine deamination to uracil than dsDNA (Frederico et al., 1990), highly vulnerable to C→T transition. These oxidative deaminations are common in cellular genomes, but are efficiently repaired (Mol et al., 1999), while such changes in ssDNA viruses might simply go unrepaired (McClelland, 1985). Alternatively, host cytidine deaminases could be increasing viral thymine content by enzymatically deaminating cytosines. These enzymes are an innate mammalian anti-viral defense, and are active against both viral RNA and ssDNA (Bishop et al., 2004). Regardless of the exact mechanism, the evidence points to a biased mutational pressure at cytosines contributing to begomovirus evolution.

Mastreviruses present a contrary case: whether or not they experience this potential thymine- enriching factor, their CUB remains unlikely in the absence of additional drivers. Mastreviruses may not experience the same mutational pressure from deamination, may have developed ways to compensate for it, or monocot hosts may interact with ssDNA genomes differently than eudicots. Neither their CP nor Rep sequences carry the signature of rapid deamination, and their CP CUB strongly adheres to host preferences, indicating the primacy of translational selection over base composition and other potential factors. Furthermore, an examination of maize streak virus revealed no evidence of the long-term C→T substitution bias evident in begomoviruses (Harkins et al., 2009). Finally, unlike most organisms that have been studied, MSV exhibits high degrees of variance in CUB between different genes, and the reasons for this variation are unclear (Adams and Antoniw, 2004). A single recently discovered eudicot-infecting mastrevirus sequence (Hadfield et al., 2012) exhibited codon usage preferences that differ from monocots, eudicots, and the other viruses we examined. Additional analysis is required to more precisely determine the forces affecting CUB in mastreviruses.

Conclusions

Codon usage bias is most often presented as the result of two competing forces: translational selection and genomic base composition. The methods most often used to evaluate it are sometimes sufficient to distinguish one of these factors from the other. However, in situations where neither factor appears significant, the available methods are of little use. Viral genomic nucleotide composition does not appear to be driving CUB in plant viruses, but there is only weak evidence of translational selection influencing CUB. Present methods are unable to explain plant virus CUB. Therefore, new ways of analyzing CUB and evaluating its likely determinants are required to more accurately parse the large amount of genomic data now available, potentially shedding light on additional factors shaping CUB, such as biased mutation rates.

Acknowledgments

Funding provided by Rutgers School of Environmental and Biological Sciences and the New Jersey Agricultural Experiment Station. We thank Kate DeRosa for assisting with data collection.

References

Adams, M.J., Antoniw, J.F., 2004. Codon usage bias amongst plant viruses. Arch. Virol. 149, 113-135.

Andret-Link, P., Fuchs, M., 2005. Transmission specificity of plant viruses by vectors. Journal of Plant Pathology 87, 153-165.

Aota, S., Ikemura, T., 1986. Diversity in G+C content at the third position of codons in vertebrate genes and its cause. Nucleic Acids Res. 14, 6345.

Au, K.G., Welsh, K., Modrich, P., 1992. Initiation of methyl-directed mismatch repair. J. Biol. Chem. 267, 12142-12148.

Bailly-Bechet, M., Danchin, A., Iqbal, M., Marsili, M., Vergassola, M., 2006. Codon Usage Domains over Bacterial Chromosomes. PLoS Comput. Biol. 2, e37.

Benevides, J.M., Stow, P.L., Ilag, L.L., Incardona, N.L., Thomas, G.J., 1991. Differences in secondary structure between packaged and unpackaged single-stranded DNA of bacteriophage phi X174 determined by Raman spectroscopy: a model for phi X174 DNA packaging. Biochemistry (Mosc.) 30, 4855-4863.

Bennetzen, J.L., Hall, B.D., 1982. Codon Selection in Yeast. J. Biol. Chem. 257, 3026-3031.

Bishop, K.N., Holmes, R.K., Sheehy, A.M., Davidson, N.O., Cho, S.-J., Malim, M.H., 2004. Cytidine Deamination of Retroviral DNA by Diverse APOBEC Proteins. Curr. Biol. 14, 1392-1396.

Bull, J.J., Molineux, I.J., Wilke, C.O., 2012. Slow Fitness Recovery in a Codon-Modified Viral Genome. Mol. Biol. Evol. 29, 2997-3004.

Cai, M.-S., Cheng, A.-C., Wang, M.-S., Zhao, L.-C., Zhu, D.-K., Luo, Q.-H., Liu, F., Chen, X.-Y., 2009. Characterization of Synonymous Codon Usage Bias in the Duck Plague Virus UL35 Gene. Intervirology 52, 266-278.

Camiolo, S., Farina, L., Porceddu, A., 2012. The Relation of codon bias to tissue-specific gene expression in *Arabidopsis thaliana*. Genetics 192, 641-649.

Cardinale, D.J., Duffy, S., 2011. Single-stranded genomic architecture constrains optimal codon usage. Bacteriophage 1, 219-224.

Chare, E.R., 2004. Selection pressures in the capsid genes of plant RNA viruses reflect mode of transmission. J. Gen. Virol. 85, 3149-3157.

Curran, J.F., Yarus, M., 1989. Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. J. Mol. Biol. 209, 65-77.

Duffy, S., Holmes, E.C., 2008. Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. J. Virol. 82, 957-965.

Duffy, S., Holmes, E.C., 2009. Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. J. Gen. Virol. 90, 1539-1547.

Frederico, L.a., Kunkel, T.a., Shaw, B.R., 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry (Mosc.) 29, 2532-2537.

Fuglsang, A., 2005. Estimating the "Effective Number of Codons": The Wright Way of Determining Codon Homozygosity Leads to Superior Estimates. Genetics 172, 1301-1307.

Gibbs, A.J., Ohshima, K., Phillips, M.J., Gibbs, M.J., 2008. The Prehistory of Potyviruses: Their Initial Radiation Was during the Dawn of Agriculture. PLoS ONE 3, e2523.

Grantham, R., Gautier, C., Guoy, M., Mercier, R., Pave, A., 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8, r49-r62.

Gray, S.M., Banerjee, N., 1999. Mechanisms of arthropod transmission of plant and animal viruses. Microbiol. Mol. Biol. Rev. 63, 128-148.

Gutierrez, C., 1999. Geminivirus DNA Replication. Cell. Mol. Life Sci. 56, 313-329.

Hadfield, J., Thomas, J.E., Schwinghamer, M.W., Kraberger, S., Stainton, D., Dayaram, A., Parry, J.N., Pande, D., Martin, D.P., Varsani, A., 2012. Molecular characterisation of dicot-infecting mastreviruses from Australia. Virus Res. 166, 13-22.

Harkins, G.W., Martin, D.P., Duffy, S., Monjane, A.L., Shepherd, D.N., Windram, O.P., Owor, B.E., Donaldson, L., van Antwerpen, T., Sayed, R.A., Flett, B., Ramusi, M., Rybicki, E.P., Peterschmitt, M., Varsani, A., 2009. Dating the origins of the maize-adapted strain of maize streak virus, MSV-A. J. Gen. Virol. 90, 3066-3074.

Hershberg, R., Petrov, D.A., 2008. Selection on codon bias. Annu. Rev. Genet. 42, 287-299.

Hiraoka, Y., Kawamata, K., Haraguchi, T., Chikashige, Y., 2009. Codon usage bias is correlated with gene expression levels in the fission yeast Schizosaccharomyces pombe. Genes Cells 14, 499-509.

Hofacker, I.L., Fekete, M., Flamm, C., Huynen, M.A., Rauscher, S., Stolorz, P.E., Stadler, P.F., 1998. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. Nucleic Acids Res. 26, 3825-3836.

Hofacker, I.L., Stadler, P.F., Stocsits, R.R., 2004. Conserved RNA secondary structures in viral genomes: a survey. Bioinformatics 20, 1495-1499.

Ikemura, T., 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. J. Mol. Biol. 151, 389-409.

Ikemura, T., 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes: Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. J. Mol. Biol. 158, 573-597.

Incardona, N.L., Prescott, B., Sargent, D., Lamba, O.P., Thomas, G.J., 1987. Phage phi X174 probed by laser Raman spectroscopy: evidence for capsid-imposed constraint on DNA secondary structure. Biochemistry (Mosc.) 26, 1532-1538.

Jenkins, G.M., Holmes, E.C., 2003. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res. 92, 1-7.

Jia, R., Cheng, A., Wang, M., Xin, H., Guo, Y., Zhu, D., Qi, X., Zhao, L., Ge, H., Chen, X., 2008. Analysis of synonymous codon usage in the UL24 gene of duck enteritis virus. Virus Genes 38, 96-103.

Jiang, Y., Deng, F., Wang, H., Hu, Z., 2008. An extensive analysis on the global codon usage pattern of baculoviruses. Arch. Virol. 153, 2273-2282.

Karlin, S., Mrázek, J., Campbell, a.M., 1998. Codon usages in different gene classes of the Escherichia coli genome. Mol. Microbiol. 29, 1341-1355.

King, A.M.Q., Adams, M.J., Carstens, E.B., Lefkowitz, E.J., 2011. Virus taxonomy: classification and nomenclature of viruses: ninth report of the international committee on the taxonomy of viruses. Elsevier Academic press, San Diego.

Krieg, A.M., Yi, A.-K., Matson, S., Waldschmidt, T.J., Bishop, G.A., Teasdale, R., Koretzky, G.A., Klinman, D.M., 1995. CpG motifs in bacterial DNA trigger direct B-cell activation. Nature 374, 546-549.

Krupovic, M., Ravantti, J.J., Bamford, D.H., 2009. Geminiviruses: a tale of a plasmid becoming a virus. BMC Evol. Biol. 9, 112.

Lefeuvre, P., Harkins, G.W., Lett, J.-M., Briddon, R.W., Chase, M.W., Moury, B., Martin, D.P., 2011. Evolutionary Time-Scale of the Begomoviruses: Evidence from Integrated Sequences in the Nicotiana Genome. PLoS ONE 6, e19193.

Liu, X., Wu, C., Chen, A.Y.H., 2010. Codon usage bias and recombination events for neuraminidase and hemagglutinin genes in Chinese isolates of influenza A virus subtype H9N2. Arch. Virol. 155, 685-693.

Liu, X.-s., Zhang, Y.-g., Fang, Y.-z., Wang, Y.-l., 2012. Patterns and influencing factor of synonymous codon usage in porcine circovirus. Virol. J. 9, 68.

Lobry, J.R., Sueoka, N., 2002. Asymmetric directional mutation pressures in bacteria. Genome Biol. 3.

Martin, D.P., Biagini, P., Lefeuvre, P., Golden, M., Roumagnac, P., Varsani, A., 2011. Recombination in Eukaryotic Single Stranded DNA Viruses. Viruses 3, 1699-1738.

McClelland, M., 1985. Selection Against dam Methylation Sites in the Genomes of DNA of Enterobacteriophages. J. Mol. Evol. 21, 317-322.

Mol, C.D., Parikh, S.S., Putname, C.D., Lo, T.P., Tainer, J.A., 1999. DNA repair machanisms for the recognition and removal of damaged DNA bases. Annual Review of Biophysics and Biomolecular Structures 28, 101-128.

Pagan, I., Holmes, E.C., 2010. Long-Term Evolution of the Luteoviridae: Time Scale and Mode of Virus Speciation. J. Virol. 84, 6177-6187.

Pedersen, S., 1984. Escherichia coli ribosomes translate in vivo with variable rate. The EMBO Journal 3, 2895-2898.

Power, A.G., 2000. Insect transmission of plant viruses: a constraint on virus variability. Current Opinions in Plant Biology 3.

Puigbò, P., Bravo, I.G., Garcia-Vallve, S., 2008. CAIcal: a combined set of tools to assess codon usage adaptation. Biol. Direct 3, 38.

Rojas, M.R., Zerbini, F.M., Allison, R.F., Robert L. Gilbertson, Lucas, W.J., 1997. Capsid Protein and Helper Component-Proteinase Function as Potyvirus Cell-to-Cell Movement Proteins. Virology 237, 283-295.

Rosario, K., Duffy, S., Breitbart, M., 2012. A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. Arch. Virol. 157, 1851-1871.

Sanjuan, R., Agudelo-Romero, P., Elena, S.F., 2009. Upper-limit mutation rate estimation for a plant RNA virus. Biol. Lett. 5, 394-396.

Sharp, P.M., Li, W.-H., 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24, 28-38.

Shen, C.K., Ikoku, a., Hearst, J.E., 1979. A specific DNA orientation in the filamentous bacteriophage fd as probed by psoralen crosslinking and electron microscopy. J. Mol. Biol. 127, 163-175.

Shields, D.C., Sharp, P.M., Higgins, D.G., Wright, F., 1988. "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. 5, 704-716.

Sun, Y., Chen, A.Y., Cheng, F., Guan, W., Johnson, F.B., Qiu, J., 2009. Molecular Characterization of Infectious Clones of the Minute Virus of Canines Reveals Unique Features of Bocaviruses. J. Virol. 83, 3956-3967.

Tsuboi, M., Tsunoda, M., Overman, S.A., Benevides, J.M., Thomas, G.J., 2010. A Structural Model for the Single-Stranded DNA Genome of Filamentous Bacteriophage Pf1. Biochemistry (Mosc.) 49, 1737-1743.

Varsani, A., Shepherd, D.N., Dent, K., Monjane, A.L., Rybicki, E.P., Martin, D.P., 2009. A highly divergent South African geminivirus species illuminates the ancient evolutionary history of this family. Virol. J. 6, 36.

Wang, L., Roossinck, M.J., 2006. Comparative analysis of expressed sequences reveals a conserved pattern of optimal codon usage in plants. Plant Mol. Biol. 61, 699-710.

Wang, M., Liu, Y.-s., Zhou, J.-h., Chen, H.-t., Ma, L.-n., Ding, Y.-z., Liu, W.-q., Gu, Y.-x., Zhang, J., 2011. Analysis of codon usage in Newcastle disease virus. Virus Genes 42, 245-253.

Wang, M., Zhang, J., Zhou, J.-h., Chen, H.-t., Ma, L.-n., Ding, Y.-z., Liu, W.-q., Liu, Y.-s., 2010. Analysis of codon usage in bovine viral diarrhea virus. Arch. Virol. 156, 153-160.

Welsh, L.C., Marvin, D.A., Perham, R.N., 1998. Analysis of X-ray diffraction from fibres of Pf1 Inovirus (filamentous bacteriophage) shows that the DNA in the virion is not highly ordered. J. Mol. Biol. 284, 1265-1271.

Wen, Z.Q., Armstrong, A., Thomas Jr, G.J., 1999. Demonstration by ultraviolet resonance Raman spectroscopy of differences in DNA organization and interactions in filamentous viruses Pf1 and fd. Biochemistry (Mosc.) 38, 3148-3156.

Woese, C.R., Kandler, O., Wheelis, M.L., 1990. Towards a natural system of organisms: Proposal for the domains of Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. U. S. A. 87, 4576-4579.

Xu, X.-z., Liu, Q.-p., Fan, L.-j., Cui, X.-f., Zhou, X.-p., 2008. Analysis of synonymous codon usage and evolution of begomoviruses. Journal of Zhejiang University SCIENCE B 9, 667-674.

Chapter 3

Mutation Spectrum of PhiX174, a Single-Stranded DNA Bacteriophage of *E. coli*

Abstract

ssDNA viruses mutate and evolve at rates comparable to RNA viruses, but through different mechanisms, since they use their hosts' high-fidelity DNA polymerases rather than low-fidelity RNA polymerases. Biased substitution patterns have been observed in ssDNA viruses, but their mutation spectrum has not been documented. Here, we generated mutation-accumulation lines of phiX174 through repeated bottleneck passages to determine base-specific mutation frequencies. We expected to see evidence of biased mutation rates consistent with oxidative damage, but our findings did not reveal any significant mutational biases.

Introduction

Unlike cellular genomes, which are exclusively double-stranded DNA (dsDNA), viral genomes can be single-stranded or double-stranded, DNA or RNA, or a combination thereof. The single-stranded DNA (ssDNA) viruses are an understudied group of viruses that are nonetheless extremely important. They are some of the most rapidly emergent viruses on earth; several ssDNA viruses have spread to novel hosts and/or geographic ranges in the late twentieth and early twenty-first centuries (Duffy and Holmes, 2007; Shackelton et al., 2005).

Only one ssDNA virus, parvovirus B19, is known to cause disease in humans, and it causes the usually mild slapped cheek rash (Servey et al., 2007), but ssDNA viruses nonetheless have a serious impact on humans. Many ssDNA viruses infect crops or livestock, and can have extremely high costs. For example maize streak virus, can severely reduce crop yields (Bosque-Pérez et al., 1998). Worryingly, these viruses often have wide host ranges (Pramesh et al., 2013). In 1997, an outbreak of a recombinant strain of East African Cassava Mosaic Virus effectively wiped out the cassava crop in Uganda, causing a severe famine (Legg and Thresh, 2000; Ndunguru et al., 2005). Despite this clear threat to global food security, relatively little work has been done to better understand how ssDNA viruses evolve and spread.

Rapid evolution is thought to be the primary driver of the ssDNA viral emergence. Long-term studies have found that ssDNA viruses evolve at approximately the same rate as RNA viruses, as measured by their substitution rates (substitutions/site/year) (Duffy et al., 2008; Sanjuan, 2012). For both RNA and ssDNA viruses, this rapid evolution is probably driven by high mutation rates, measured as mutations/site/replication. Mutation rates in ssDNA viruses more closely resemble RNA viruses than the slower-evolving and slower-mutating dsDNA viruses (Duffy et al., 2008). How these viruses achieve high mutation rates is not clear.

High RNA mutation rates are driven by their use of lower-fidelity RNA-dependent RNA polymerase (RdRp) for genome replication. Most viral RdRp lacks the 3’→5’ exonuclease ability of high-fidelity DNA polymerases, so replication errors persist at higher rates compared to DNA replication (Steinhauer and Holland, 1986; Ward et al., 1988). However, all ssDNA viruses utilize their hosts' high-fidelity DNA polymerases, which have 3'→5' exonuclease activity, so this mechanism cannot explain the high mutation rates of ssDNA viruses.

Long-term evolution studies have provided a possible explanation. ssDNA viruses exhibit a persistent substitution bias from cytosine to thymine (Duffy and Holmes, 2009). This bias may be due to spontaneous deamination of cytosine to uracil. If uncorrected, deamination results in a transition from C to T following replication. Unpaired cytosine is more susceptible to spontaneous deamination than base-paired cytosine (Frederico et al., 1990), possibly because the amino group is stabilized by hydrogen bonding when paired with guanine. While the mutation rate of ssDNA viruses have been documented (Cuevas et al., 2009; Drake, 1991), and experimental evolution studies have charted the evolutionary trajectories ssDNA phages operating under selection (Dickins and Nekrutenko, 2009), the spontaneous mutation spectrum of ssDNA viruses under neutral conditions has not yet been determined. The beneficial mutational spectrum assessed through the short-term evolution of phiX174, a ssDNA bacteriophage, is consistent with the mutations most likely to occur due to deamination and other spontaneous oxidative reactions (Rokyta et al., 2005).

We used plaque-to-plaque bottleneck passaging to evolve bacteriophage populations under neutral conditions, to allow for mutation accumulation in the absence of selection against deleterious mutations. Under plaque-to-plaque passaging, phage populations pass through a bottleneck with a population size of one during each passage, as each individual plaque is founded by a single phage, and each subsequent passage is derived from an arbitrarily chosen single plaque of the previous passage. Such strict bottlenecks sharply reduce diversity (Li and Roossinck, 2004), allowing for fixation of non-lethal mutations. Repeated bottleneck passages therefore can facilitate mutation accumulation, providing a picture of the mutation spectrum of the virus population.

We allowed ten populations of phiX174, each originating from an independent plaque, to evolve under bottlenecking for 50 passages on its host, *Escherichia coli*, to facilitate mutation accumulation. A single plaque from each population at the end of the experiment was sequenced to determine the mutational spectrum of this ssDNA virus.

Methods

Preparation of PhiX174 Populations

Ten independent populations of bacteriophage phiX174 were established from a wild-type ancestral population. 10µl from the ancestral phiX174 population was used to seed ten TK agar plates along with 100µl *E. coli* C122 in 3ml TK top agar (0.7% concentration, soft agar). These plates were incubated at 37˚C overnight. The following day, the top agar was scraped and collected. Each plate was then washed with 3ml liquid TK media to collect additional phage, which was combined with the top agar. These mixtures were centrifuged at 3000 rpm for 10 minutes to pellet bacterial cells, and the supernatant from each mixture was filtered through a 0.22µm filter to remove cellular matter, resulting in ten purified lysates of phiX174.

Serial Bottleneck Passaging

Bottleneck passaging was used to allow drift to overwhelm selection as the phage populations evolved. Fifty single-plaque transfers were used to repeatedly force each phage population through fifty bottlenecks with a population size of one. To initiate passaging, a sample from each starting population was diluted to a concentration of approximately 10^3 plaque forming units (PFU)/ml. From this, 100µl was plated on TK plates with 100µl *E. coli* from a stationary phase overnight culture in 3ml TK top agar, so that there were approximately 10-100 plaque forming units (PFU) per plate, to prevent the resulting plaques from touching or overlapping, which would have increased the potential genetic diversity within such plaques. After phage were plated, plates were incubated at 37˚C for three to five hours, until plaques were just visible against the growing *E. coli* lawn. Once plaques were evident, the single plaque closest to an arbitrary mark made on the underside of each plate prior to incubation was chosen. This plaque was gently touched with a 10µl pipette tip, collecting some of the phage from that plaque, and the tip was briefly shaken in 1ml of liquid TK media to dilute the phage, and that dilution used to initiate the next passage as described above. This process was repeated 50 times. Samples of phage from every fifth passage, as well as initial and final phage populations, were preserved in glycerol at -80˚C. The total time of the growth stage of all fifty passages was 178 hours, an average of ~3.5 hours of incubation per passage. Conservatively assuming exponential *E. coli* growth for only the last hour of lawn formation and two generations of phage replication per hour in exponentially-dividing *E. coli*, this corresponds to 100 generations of phage replication.

Sequencing

Sanger sequencing was used to determine the sequences of ancestral and final populations. PhiX174 genomes were amplified using PCR with XL (extra-long) Taq (Applied Biosystems) which allowed for amplification of the 5.4kb phiX174 genome in three reactions. PCR products were verified through gel electrophoresis on a 0.8% agarose gel run at 100V for 30 minutes, and subsequently purified using ExoSAP (Affymetrix). Purified DNA was sequenced off-site using Sanger methods by GeneWiz.

The resulting sequences were manually checked for errors and automatically aligned to the ancestral genome with Sequencher v4.10 (Gene Codes Corporation). Alignments were also manually checked for errors. Pairwise comparisons between ancestral and derived sequences for each line were made, and the frequency of each nucleotide substitution determined. All mutations were confirmed by at least two separate sequencing reactions. Mutation rates were calculated as the number of mutations per site per generation across all fifty passages.

Results

After 50 passages, a total of six unique mutations appeared across the ten evolved lines; each line exhibited zero to two mutations. Mutation rates for lines in which mutations occurred ranged from 1.86x10^{-6} mutations/site/generation (m/s/g) to 3.71x10^{-6} m/s/g. Only three distinct substitutions were observed (A→G, C→T, T→C, Table 1). Unexpectedly, T→C transitions were the most common. Three observed mutations (A1889G, A2061G, and T3235C) were nonsynonymous, while three (T1012C, C1123T, and T3245C) were synonymous (Table 1).

Table 1. Observed mutations, the region in which they occur, and the effects on protein sequence. Grey indicates synonymous changes.

ORF	Mutation	Amino Acid
F	T1012C	I4I
	C1123T	S41S
	A1889G	T297A
	A2061G	Y354C
H	T3235C	L102S
	T3245C	L105L

Discussion

Before implementing bottleneck passaging and full genome sequencing, we attempted several alternative approaches to determine the mutation spectrum of phiX174. The initial objective was to utilize gene F or H amber mutant strains of phiX174 grown on trans-complimenting *E. coli* hosts (those that express F or H off of a plasmid). This would have freed up a large region within the phiX174 genome to evolve neutrally while its gene function was provided by plasmid copy. However, all of the amber mutant strains of phiX174 we obtained from Bentley Fane proved to be unstable over relatively few generations, so were unsuitable for long-term mutation accumulation.

Our second approach was to utilize the B-free strain of phiX174, which has lost the functionality of gene B, the internal scaffolding protein, also created by Ben Fane (Chen et al., 2007). Use of this would have provided a region of the genome under relaxed selection relative to wild-type phiX174, providing a basis for a comparison between neutral and selected evolution. However, gene B is entirely overlapping with gene A/A*, in an alternative reading frame, so the relaxed selection on that region would only have applied to synonymous sites within the gene A reading frame, not the entirety of gene B, making this a less attractive option than the amber mutants initially attempted. We attempted to use B-free phiX174, but this strain suffers from slower and less robust growth compared to wild-type phiX174. Consequently, the plaques formed were too small to reliably pick for plaque-to-plaque transfers.

For these reasons, we settled on bottleneck passaging of wild type phage followed by full-genome sequencing to characterize the mutation spectrum of phiX174. Bottleneck passaging has been used to overwhelm selection with genetic drift in populations of many viruses. Repeated bottleneck passages are often associated with significant decreases in fitness, demonstrating Muller's Ratchet as deleterious mutations accumulate (Burch and Chao, 1999; Clarke et al., 1993; Novella, 2004). It has also been used to study the mutations that accumulate in population in the absence of selection (Li and Roossinck, 2004; Yuste et al., 2000).

ssDNA viruses evolve at rates comparable to RNA viruses (Sanjuan, 2012), and while the mechanisms are not understood, it is probably due to high intrinsic mutation rates. Unlike RNA viruses, though, these high rates are unlikely to be due to polymerase errors, since ssDNA viruses utilize their hosts' high-fidelity polymerases. Long-term studies of ssDNA virus evolution have provided evidence of a persistent C→T substitution bias, which may be explained by an elevated rate of C→T mutations due to oxidative degradation of unpaired ssDNA bases (Duffy and Holmes, 2008; Duffy and Holmes, 2009).

However, our experimentally derived mutational spectrum of phiX174 do not support the overrepresentation of oxidation-induced mutations relative to others. We simply have too few mutations to draw any conclusions about the mutation spectrum of phiX174 – fewer than one mutation per bottlenecked lineage – and those we did observe are not indicative of oxidative damage to ssDNA genomes. Many additional generations of mutation accumulation may have provided a clearer picture of the mutation spectrum of phiX174. Other experiments using phiX174 or an RNA phage with a comparable measured mutation rate (Chao et al., 2002) have found much more rapid mutation accumulation (Dennehy et al., 2013; Rokyta et al., 2005).

Our data offer no opposition to the null hypothesis that polymerase error is the primary source of mutations in phiX174. The mutation rates we observed were comparable to the previously measured phiX174 mutation rate (Cuevas et al., 2009), and slower than those measured in another ssDNA bacteriophage, M13 (Drake, 1991). In contrast, the error rate of *E. coli* polymerase III has been estimated to be approximately 10^{-7} – 10^{-8} mutations/base/replication (Fijalkowska et al., 2012), which is somewhat slower than the mutation rates we observed. *E. coli* post-replication mismatch repair corrects the majority of these errors, bringing its overall mutation rate down to $5x10^{-10}$ mutation/site/replication

(Fijalkowska et al., 2012). The major repair mechanism is based on recognition of methylated GATC tetranucleotides in the template strand of a newly replicated chromosome, and efficiently repairs replication-induced mismatches not caught by polymerase III 3'-5' exonuclease activity (Marti et al., 2002). PhiX174 might be expected to experience slower mutation rates due to this repair mechanism, but the phiX174 strain we used contains zero GATC sites instead of the expected five to six (Cuevas et al., 2011). Consequently, it cannot be subject to post-replication repair via the *E. coli* methyl-directed mismatch repair system, a conclusion that is supported by the observation that artificial inclusion of GATC sites into the phiX174 genome can significantly decrease its mutation rate (Cuevas et al., 2011). For these reasons, mutation rates driven by polymerase induced errors in phiX174 would be expected to resemble the error rates of the *E. coli* polymerase III holoenzyme, rather than *E. coli* as a whole.

The specific mutations we observed do not match the profile of expected errors due to spontaneous oxidation of ssDNA. Oxidation can drive several mutations, most rapidly C→T and A→G through oxidative deamination, and G→T through an 8-oxoguanine intermediate. In a test of the one-step adaptive landscape of phiX174, each of the 20 observed beneficial mutations sequenced were one of these three, strongly indicative of mutation rate augmented or driven by oxidative damage (Rokyta et al., 2005). In contrast, just half (3/6) of our observed mutations were consistent with oxidation, indicating that it was not the primary driver. Importantly, three of the six mutations found were T→C mutations, the exact opposite substitution of the most common kind of oxidative mutation, and one that is highly underrepresented in long term phiX174 evolution (Yee Mey Seah, personal communication).

There is a well-documented mutational bias in which transitions are more frequent than transversions (Gojobori et al., 1982; Yang and Yoder, 1999), and our observations are in line with these previous findings. All six of the mutations we observed were transitions. This bias may be due to the kinetics of DNA replication, in which the likelihood of a misincorporated base remaining is a product of the speed with which the next base is inserted compared to the speed with which the misincorporated base is removed via 3'-5' exonuclease activity (Kunkel and Bebenek, 2000). Since transversions distort the double helix more than transitions, they are far more likely to be excised rather than extended, while transitions may result in milder distortions, allowing for extension despite the mispairing (Kunkel and Bebenek, 2000). Therefore, the mutations we observed are most consistent with polymerase error during DNA replication.

More mutations may have accumulated in a large, non-coding region; half of the mutations documented in Dennehy et al. (2013) were found in a very small non-coding region of an RNA phage genome. PhiX174 has extremely small intergenic regions, limiting this possibility. The length of the growth phase between bottlenecks may also have affected mutation accumulation. The relationship between this growth phase and the mutation rate of the subject organism are two factors that can strongly effect the rate at which mutations accumulate (Manrubia et al., 2005). Given that phiX174 experiences mutations at rates comparable to a dsRNA virus but slightly slower than ssRNA viruses

(Sanjuan, 2012), this period may have been too short, preventing mutations from accumulating at the expected rate. More passages or longer growth periods may overcome these difficulties and yield more informative results.

References

Bosque-Pérez, N.A., Olojede, S.O., Buddenhagen, I.W., 1998. Effect of maize streak virus disease on the growth and yield of maize as influenced by varietal resistance levels and plant stage at time of challenge. Euphytica 101, 307-317.

Burch, C.L., Chao, L., 1999. Evolution by Small Steps and Rugged Landscapes in the RNA Virus œï6. Genetics 151, 921-927.

Chao, L., Rang, C.U., Wong, L.E., 2002. Distribution of Spontaneous Mutants and Inferences about the Replication Mode of the RNA Bacteriophage œÜ6. J. Virol. 76, 3276-3281.

Chen, M., Uchiyama, A., Fane, B.A., 2007. Eliminating the Requirement of an Essential Gene Product in an Already Very Small Virus: Scaffolding Protein B-free øX174, B-free. J. Mol. Biol. 373, 308-314.

Clarke, D.K., Duarte, E.A., Moya, A., Elena, S.F., Domingo, E., Holland, J., 1993. Genetic Bottlenecks and Population Passages Cause Profound Fitness Differences in RNA Viruses. J. Virol. 67, 222-228.

Cuevas, J.M., Duffy, S., Sanjuan, R., 2009. Point Mutation Rate of Bacteriophage X174. Genetics 183, 747-749.

Cuevas, J.M., Pereira-Gómez, M., Sanjuán, R., 2011. Mutation rate of bacteriophage ΦX174 modified through changes in GATC sequence context. Infect. Genet. Evol. 11, 1820-1822.

Dennehy, J.J., Duffy, S., O‚ÄôKeefe, K.J., Edwards, S.V., Turner, P.E., 2013. Frequent Coinfection Reduces RNA Virus Population Genetic Diversity. J. Hered.

Dickins, B., Nekrutenko, A., 2009. High-Resolution Mapping of Evolutionary Trajectories in a Phage. Genome Biol. Evol. 1, 294-307.

Drake, J.W., 1991. A constant rate of spontaneous mutation in DNA-based microbes. Proc. Natl. Acad. Sci. U. S. A. 88, 7160-7164.

Duffy, S., Holmes, E.C., 2007. Multiple Introductions of the Old World Begomovirus Tomato yellow leaf curl virus into the New World. Appl. Environ. Microbiol. 73, 7114-7117.

Duffy, S., Holmes, E.C., 2008. Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. J. Virol. 82, 957-965.

Duffy, S., Holmes, E.C., 2009. Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. J. Gen. Virol. 90, 1539-1547.

Duffy, S., Shackelton, L.A., Holmes, E.C., 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. 9, 267-276.

Fijalkowska, I.J., Schaaper, R.M., Jonczyk, P., 2012. DNA replication fidelity in Escherichia coli: a multi-DNA polymerase affair. FEMS Microbiol. Rev. 36, 1105-1121.

Frederico, L.a., Kunkel, T.a., Shaw, B.R., 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry (Mosc.) 29, 2532-2537.

Gojobori, T., Li, W.-H., Graur, D., 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. 18, 360-369.

Kunkel, T.A., Bebenek, K., 2000. DNA Replication Fidelity. Annu. Rev. Biochem. 69, 497-529.

Legg, J.P., Thresh, J.M., 2000. Cassava mosaic virus disease in East Africa: a dynamic disease in a changing environment. Virus Res. 71, 135-149.

Li, H., Roossinck, M.J., 2004. Genetic Bottlenecks Reduce Population Variation in an Experimental RNA Virus Population. J. Virol. 78, 10582-10587.

Manrubia, S.C., Escarmís, C., Domingo, E., Lázaro, E., 2005. High mutation rates, bottlenecks, and robustness of RNA viral quasispecies. Gene 347, 273-282.

Marti, T.M., Kunz, C., Fleck, O., 2002. DNA Mismatch Repair and Mutation Avoidance Pathways. J. Cell. Physiol. 191, 28-41.

Ndunguru, J., Legg, J.P., Aveling, T.A.S., Thompson, G., Fauquet, C.M., 2005. Molecular biodiversity of cassava begomoviruses in Tanzania: evolution of cassava geminiviruses in Africa and evidence for East Africa being a center of diversity of cassava geminiviruses. Virol. J. 2, 21.

Novella, I.S., 2004. Negative Effect of Genetic Bottlenecks on the Adaptability of Vesicular Stomatitis Virus. J. Mol. Biol. 336, 61-67.

Pramesh, D., Mandal, B., Phaneendra, C., Muniyappa, V., 2013. Host range and genetic diversity of croton yellow vein mosaic virus, a weed-infecting monopartite begomovirus causing leaf curl disease in tomato. Arch. Virol. 158, 531-542.

Rokyta, D.R., Joyce, P., Caudle, S.B., Wichman, H.A., 2005. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. Nat. Genet. 37, 441-444.

Sanjuan, R., 2012. From molecular genetics to phylodynamics: evolutionary relevance of mutation rates across viruses. PLoS Pathog. 8.

Servey, J.T., Reamy, B.V., Hodge, J., 2007. Clinical Presentations of Parvovirus B19 Infection. Am. Fam. Physician 75, 373-376.

Shackelton, L.a., Parrish, C.R., Truyen, U., Holmes, E.C., 2005. High rate of viral evolution associated with the emergence of carnivore parvovirus. Proc. Natl. Acad. Sci. U. S. A. 102, 379-384.

Steinhauer, D.A., Holland, J.J., 1986. Direct method for quantitation of extreme polymerase error frequencies at selected single base sites in viral RNA. J. Virol. 57, 219-228.

Ward, C.D., Stokes, M.A., Flanegan, J.B., 1988. Direct measurement of the poliovirus RNA polymerase error frequency in vitro. J. Virol. 62, 558-562.

Yang, Z., Yoder, A.D., 1999. Estimation of the Transition/Transversion Rate Bias and Species Sam. J. Mol. Evol. 48, 272-283.

Yuste, E., López-Galíndez, C., Domingo, E., 2000. Unusual Distribution of Mutations Associated with Serial Bottleneck Passages of Human Immunodeficiency Virus Type 1. J. Virol. 74, 9546-9552.

Chapter 4

Sodium Bisulfite as a Treatment Against ssDNA Viruses

Abstract

Lethal mutagenesis is the use of a mutagen to push a population of pathogens to extinction through the accumulation of deleterious mutations. This has been considered as a potential treatment for viral infections, especially those caused by viruses that experience high mutation rates that make effective treatment difficult. Although the theory is sound, and mutagens are used to treat certain viral infections, the efficacy of this approach is difficult to confirm; other factors may be at work during mutagenic treatment, and efforts to document the negative effects of mutagenesis in the laboratory have sometimes resulted in fitness increases in the target populations. Here, we used a cytosine-specific mutagen to drive populations of phiX174, a single-stranded DNA bacteriophage, to extinction. Since we have eliminated many alternative factors that could have contributed to the fitness collapse of treated populations, mutagenesis is a probable contributor to our results, though it is unlikely to be the sole driver of viral extinction.

Introduction

Infections caused by single-stranded DNA (ssDNA) viruses have a high economic and human cost. Though there are no severe human diseases caused by ssDNA viruses, they are important pathogens of crops and livestock. For example, porcine circoviruses causes millions of Euros of losses to the pork industry in Europe annually (Armstrong and Bishop, 2004). ssDNA crop pathogens are even more costly, inflicting not only economic, but also human, costs. A 1997 outbreak of a novel recombinant form of East African Cassava Mosaic Virus caused a severe famine in Uganda (Legg and Fauquet, 2004).

Although only one characterized ssDNA virus is a human pathogen (erythrovirus B19, which causes slapped cheek rash, a usually mild disease of children (Servey et al., 2007)), they are emergent threats. In the last few decades, a number of novel animal pathogens have emerged as the result of a ssDNA virus shifting to a new host. Canine Parvovirus 2 and Canine Parvovirus 2a (CPV2a), which both evolved from feline panleukopenia virus, did not exist until the late 1970s, and spread around the world by 1990 (Parrish and Kawaoka, 2005).

Vaccines are frequently employed against many types of viruses, single-stranded or double-stranded, DNA or RNA. Some confer lifetime immunity, while other do not owing to the rapidity with which the target virus evolves (for example, influenza, (Buonagurio et al., 1986)). In other cases, rapid evolution of the target virus precludes vaccination (for example, HIV (Ball et al., 2007)) . In addition to vaccines, drugs are available to combat many viruses. Tamiflu, for example, competitevly binds to receptors used by the influenza virus to infect human cells (Ward, 2005). Ribavirin, an RNA nucleoside analog, has been successfully employed against several RNA viruses (Bodenheimer et al., 1997; Crotty, 2001), though the precise mechanism is unclear.

Despite their rapid emergence and ability to inflict high economic and human costs, few of these measures are available against ssDNA viruses. Some animals are routinely vaccinated against common infections, such as CPV2a (Yule et al., 1997), but there are no drugs available for ssDNA viral infections. Additionally, the rapidity with which ssDNA viruses evolve may preclude effective vaccination. Short-term mutation rates in ssDNA viruses are comparable to those of RNA viruses, rather than those of other DNA-based organisms (Drake, 1991; Raney et al., 2004; Sanjuán et al., 2010). RNA viruses experience rapid mutations because RNA-dependent RNA polymerases, used by all RNA viruses for genome replication, lack the error-checking function present in processive, high-fidelity DNA polymerases (Steinhauer and Holland, 1986; Ward et al., 1988). Consequently, the inherent error rate associated with genome replication in RNA viruses is significantly higher than that of double-stranding DNA viruses or cellular organisms (Kunkel, 2004; Steinhauer and Holland, 1986). All ssDNA viruses use their hosts DNA polymerases, which should imply an extremely low polymerase-induced error rate.

ssDNA viruses have been documented to avoid host DNA repair pathways, which would increase their mutation rate relative to that of their dsDNA host cells; phiX174 lacks the requisite GATC sites to stimulate the *Escherichia coli* methyl-directed mismatch repair system (McClelland, 1985), and when artificially modified to contain more GATC tetranucleotides, experienced a mutation rate one to two orders of magnitude lower than the unmodified phage (Cuevas et al., 2011). However, this difference was insufficient to completely explain the total difference in mutation rates between the ssDNA phage and their hosts, and there must be additional sources of mutations affecting ssDNA viral genomes. The most likely candidate would be the inherent instability of cytosine in single-stranded DNA. Cytosine is the most chemically unstable nucleotide used in DNA or RNA. It is particularly prone to spontaneous deamination, in which the pyrimidines ring is oxidized and its amino group is lost in the form of ammonia and replaced with a ketone, leaving uracil. If unrepaired, this uracil base pairs with adenine during replication, which results in a C to T transition. Methylcytosine, commonly found in DNA, deaminates directly to thymine. The amino group attached to the 4' carbon may be stabilized during base pairing with guanine, as spontaneous deamination is 100 times more likely in unpaired cytosine (Frederico et al., 1990).

No one has determined the base-specific mutation rates of any ssDNA virus, so it is not possible to say with certainty that rapid and spontaneous cytosine deamination accounts for the observed rates of mutations and substitutions in ssDNA viruses. However, evidence from codon usage bias studies indicate that rapid cytosine mutations are likely: T-ending codons tend to be overrepresented in many ssDNA viruses, which would be expected if cytosine experiences rapid spontaneous deamination (Cardinale et al., 2013; Cardinale and Duffy, 2011). C to T transitions might be negatively selected if they occur at non-synonymous sites, but tolerated when amino acid sequences are conserved.

This may indicate a vulnerability that can be exploited to treat ssDNA viruses: further elevation of the cytosine mutation rate may induce lethal mutagenesis. Lethal mutagenesis is an antiviral therapy based in population genetics: deleterious mutations are far more common than beneficial ones, so at some threshold mutation rate, the negative effects of accumulating mutations will decrease the fitness of individuals, on average, below replacement. At the population level, if each virus is unable to make, on average, at least one viable progeny, the population will eventually go extinct (Bull et al., 2007).

Mutagenic treatment has been used against RNA viruses. Experimentally, mutagens can be useful against a wide range of viruses, such as bacteriophages, influenza, and HIV (Hayatsu and Miura, 1970; Pauly et al., 2015; Vivet-Boudou et al., 2015). Ribavirin is a nucleoside analog that has been successfully employed in vitro to treat Hepatitis C (Bodenheimer et al., 1997; Crotty, 2001; Mangia et al., 2005). Some evidence suggests that it works by elevating the mutation rate in the target viruses, pushing them past the lethal mutagenesis threshold, though this is not a universally accepted explanation (Graci and Cameron, 2008). In addition to inducing mutations, mutagens can affect intracellular signaling, translation, genome replication, and host fitness, each of which can impact viral replication and viability (Bull et al., 2013; Graci and Cameron, 2006). Despite the uncertainty, it is likely that mutagenesis plays at least some role in the ability of nucleoside analogs like ribavirin to combat RNA viruses (Graci and Cameron, 2006).

Lethal mutagenesis is an attractive option to treat RNA viruses, but is typically not viable against DNA viruses. RNA viruses tend to mutate much faster than DNA viruses (Duffy et al., 2008), meaning they are much closer to the lethal mutagenesis threshold, requiring just a small nudge to go extinct. Consequently, lower doses of mutagen are required, which makes treatment less likely to adversely affect the host cell or organism. Conversely, DNA viruses tend to mutate more slowly, so a higher concentration of the mutagen would be required to reach the threshold. Furthermore, unlike RNA base analogs, DNA mutagens will affect the host genome as well as the target viral genome.

Despite these challenges, ssDNA viruses may be uniquely susceptible to cytosine mutagenesis. ssDNA viruses experience mutations as rapidly as many RNA viruses (Duffy et al., 2008), so lower doses of mutagen should be required for treatment, minimizing collateral damage to the host. More importantly, since it is likely that the cytosine-specific mutation rate in ssDNA viruses is extremely high, cytosine can be targeted specifically for mutagenesis, ignoring the slower-mutating bases. Since spontaneous deamination is probably responsible for elevated cytosine mutation rates in ssDNA viruses (Frederico et al., 1990), a deaminating agent would be the ideal candidate.

Furthermore, cells tend to efficiently repair spontaneous cytosine deamination in their genomic DNA. Cells have mechanisms to effectively reverse the effects of cytosine deamination in their own genomes, specifically the Uracil N-glycosylase enzyme (UNG) which excises uracil from DNA leaving an abasic site (Marti et al., 2002). Mammalian cells are constantly repairing cytosine deamination in the coding strand of highly expressed genes, which are often single-stranded and not associated with transcription machinery (Majewski, 2003). The prevalence of this repair enzyme throughout cellular life lends credence to the theory that unpaired cytosine is significantly more susceptible to spontaneous deamination than cytosine base-paired with guanine. Therefore, if cytosine could be specifically targeted with a deaminating agent, ssDNA viruses, which appear to not be affected by UNG, may be highly susceptible, while their hosts may be largely immune to the treatment.

Sodium bisulfite ($NaHSO_3$) is a candidate mutagen for this treatment. It is a deaminating agent that specifically targets cytosine. It is a common preservative and food additive, and has a Generally Recognized as Safe (GRAS) certification from the FDA (Nair et al., 2003). It has been documented as a mutagen in viruses, bacteria, and yeast (Hayatsu and Miura, 1970; Schimz, 1980), but not in multicellular eukaryotes (Nair et al., 2003). We treated populations of a ssDNA bacteriophage with sodium bisulfite and evaluated their viability after up to four weeks of exposure. We also measured growth rate in treated an untreated phage, and sequenced treated phage to determine the genetic basis for any changes in fitness that were observed.

Methods

Phage preparation

PhiX174 stock solutions were generated from wild-type phiX174 generously provided by Dr. Bentley Fane. Stocks were diluted to approximately 10^5 PFU/ml and plated on 100uL stationary *E. coli* C122 (also provided by Ben Fane) in 3 ml TK top agar (0.7% concentration, soft agar) on TK plates (Fane and Hayashi, 1991). Plates with phage and *E. coli* were incubated overnight at 37°C.

The following day, the top agar of each plate was scraped with a sterilized spatula and separately collected in four 15ml tubes. Each plate was then washed with 3 ml liquid TK, which was also collected. These tubes were centrifuged at 3000rpm for 10 minutes. Following centrifugation, the supernatant was filtered through a 0.22μm filter. The resulting purified phage were stored at 4°C. Four independent populations were generated in this manner.

Bisulfite Treatment

Solutions of liquid TK media with 0M, 0.2M, and 0.4M sodium bisulfite, all containing 0.5M sodium acetate, were prepared from a 4M stock of sodium bisulfite and a 3M stock of sodium acetate to a volume of 5.0ml. Four 0.2M and 0.4M treatments were prepared, each with an accompanying untreated control. 10μl from each of the four purified phiX174 lysates was added to treatment and control tubes, which facilitated pairwise comparison between each treatment condition and the same phage population in a control environment. These populations were incubated at room temperature for the duration of the experiment.

Immediately following the addition of phage, each tube was sampled to determine phage concentration. 10μl from each tube was plated on TK plates with 3ml TK top agar and 100μl stationary *E. coli* C122. These plates were incubated overnight at 37°C. The following day, plaques were counted to determine phage concentration in each treated or control population. Each population was sampled to determine viable phage concentration every 2 hours for the first 12 hours of treatment, then again 24 hours after the start of treatment, every 24 hours thereafter until 14 days elapsed, and every 48 hours thereafter until no viable phage were evident. The maximum duration of the experiment was 28 days for the 0.2M treatment, and 11 days for the 0.4M treatment.

These protocols were repeated with bacteriophage T4, to assess the effects of sodium bisulfite on the viability of a phage with a dsDNA genome. The protocol was exactly the same as for phiX174, except that each T4 lineage was carried out to the maximum duration of phiX174 sampling for its level of treatment (28 days for 0.2M, 11 days for 0.4M). T4 viability was determined through the same protocols and at the same intervals as phiX174 viability.

Burst Time

Burst time assays were carried out for three independent phiX174 populations treated with 0.2M sodium bisulfite. 0.2M bisulfite treatments were prepared as described above. 100µl of purified phiX174 lysate from three independently derived populations was added to each treatment, and phage populations were tested 24 and 48 hours following initial exposure. To determine phage burst times, liquid cultures of *E. coli* were grown overnight to stationary phase. Each day, 100µl of an overnight culture of *E. coli* was added to 5ml liquid TK media, which was incubated at 37°C, shaking at 110 rpm for 3 hours, to reach exponential phase. 100µl of each treatment phage population was then added to an exponentially growing *E. coli* culture, for an initial phage concentration of approximately 10^3 PFU/ml and initial cell concentration of approximately 10^5 – 10^6 CFU/ml. These were incubated for 1 minute at room temperature to allow for phage attachment. After 1 minute, tubes were centrifuged at 3000 rpm for 10 minutes to pellet cells with attached phage, while leaving unattached phage in the supernatant. The supernatant was discarded, and cells were resuspended in 5ml liquid TK media. Resuspension denoted the start of the assay.

Immediately following resuspension, 100μl from each population was plated on 100μl of an overnight culture of *E. coli* in 3ml TK top agar on TK plates. Resuspended bacterial populations were then incubated at 37°C shaking at 110 rpm. Ten minutes after resuspension, phage populations were again sampled and plated. Sampling and plating continued every 2 minutes until 60 minutes had elapsed from the time of resuspension. Plates we incubated overnight at 37°C and plaques were counted the following day to determine phage concentration over time. Burst time was determined by the point at which the bacteriophage populations began to increase, as measured by PFU/ml.

Genome Sequencing

Solutions of 0M, 0.2M, and 0.4M sodium bisulfite were prepared as described above. 100μl purified phiX174 lysate was added to each to initiate treatment. After 14 days, samples from each population were provided to the Rutgers SEBS Genome Cooperative, where libraries were prepared using the Illumina TruSeq RNA Library Prep Kit v2 rather than a dsDNA kit, treating the phiX174 ssDNA as first strand cDNA. This was done because there is no kit for ssDNA replication, and the standard method of amplifying ssDNA via rolling circle replication (RCR) with Templiphi strongly biased reads towards the phiX174 origin of replication, since phiX174 naturally uses RCR during its replication. Libraries were verified using an Agilent Bioanalyzer High Sensitivity DNA chip. These libraries were subjected to MiSeq Illumina sequencing.

Resulting reads were assembled to the phiX174 reference genome (GenBank accession NC_001422). Following assembly, the percent of reads of each base at each site were computed. Sites with fewer than 10,000 reads were excluded from further analysis to minimize the effects of Illumina errors.

Effect on the host *E. coli*

Bisulfite treated *E. coli* were evaluated to determine if the mutagen adversely affected the growth rate or viability of the bacterial hosts, which could potentially influence phage viability. If *E. coli* was adversely affected by bisulfite treatment, that could explain some of the differences in viability between treated and untreated phage.

To determine the effects of bisulfite treatment on *E. coli* viability, we prepared liquid cultures of *E. coli* as described above. These were diluted to approximately 10^7 CFU (colony forming units)/ml in solutions containing 0.02M or 0.04M sodium bisulfite and 0.05M sodium acetate. These concentrations were used instead of the tenfold higher concentrations used to treat phage because there was a minimum tenfold dilution between the treated phage cultures and plating the phage on bacterial hosts, so *E. coli* were never exposed to bisulfite concentrations higher than 0.04M over the course of the experiment. Cells were diluted to approximately 10^4/ml spread on TK plates, and incubated at 37°C overnight. The following day, colonies were counted to determine the concentration of viable bacteria in treated and untreated cultures.

To further evaluate the effects on *E. coli* viability during growth in top agar, *E. coli* lawns were plated exactly as described above during determination of phage viability for treated and untreated phage, except that no phage were present in the 0M, 0.2M, and 0.4M bisulfite solutions. These plates were also incubated overnight at 37°C. The following day, they were removed and the quality o*f* resulting *E. coli* lawns was evaluated.

Finally, we carried out growth curves of untreated, 0.02M bisulfite treated, and 0.04M bisulfite treated *E. coli*. Three liquid cultures of *E. coli*, one colony in 25 ml liquid TK media, were prepared and incubated overnight at 37°C, shaking at 110 rpm. The following day, 10µl of each overnight culture was added to 5 ml solutions of TK media and 0.05M sodium acetate, with either 0M, 0.02M, or 0.04M sodium bisulfite. These solutions were sampled immediately for *E. coli* concentration by spreading 10µl on a TK plate. They were then incubated at 37°C shaking at 110 rpm. Every hour, each culture was sampled for *E. coli* concentration. After ten hours, TK plates with *E. coli* samples were incubated at 37°C overnight. The following day, the colonies on these plates were counted to determine the concentration of *E. coli* in treated and untreated cultures over the course of the ten hour sampling period.

Transmission Electron Microscopy

Control and treatment phage populations were generated for 0.2M and 0.4M sodium bisulfite treatment, as described above. Populations were incubated at 25°C for 14 days. Following incubation, each population was negative stained on carbon grids at the Robert Wood Johnson Medical School Department of Pathology Core Imaging Lab. After drying, each sample was subjected to transmission electron microscopy at 75,000x and 100,000x magnification using

Philips CM12 transmission electron microscope to determine if bisulfite treatment impacted capsid structure.

Results

Phage Viability

All lines of treated and untreated phage initially contained approximately 10^8 PFU/ml. Treatment of phiX174 virions with 0.2M sodium bisulfite resulted in complete loss of detectable viability (<10 PFU/ml) in 28 days (672 hours, Figure 1A). After 24 days (576 hours), no viability was detectable in two of the treated lines, and by day 28 no viable phage were detectable in any of the treated lines. Control (untreated) lines maintained a near-constant concentration, approximately 10^7-10^8 PFU/ml for the duration of the 28-day treatment period (Figure 1A).

At the start of 0.4M sodium bisulfite treatment, all phage lines exhibited concentrations of approximately 10^8 PFU/ml. One treated line fell below the detection threshold after seven days, while a second did so after nine days, and the final line exhibited no detectable viability after 11 days of bisulfite exposure (Figure 1B). Throughout this time, untreated lines exhibited unchanged concentrations of approximately 10^8 PFU/ml.

At the onset of 0.2M sodium bisulfite treatment of bacteriophage T4, all lines exhibited phage concentrations of 10^7-10^8 PFU/ml. All four untreated lines still showed approximately 10^7 PFU/ml after 28 days, while phage density in the four treated lines varied from approximately 10^4 to 10^6 PFU/ml (Figure 2A).

All T4 lines, treated and untreated, exhibited 10^7-10^8 PFU/ml at the start of 0.4M sodium bisulfite treatment. After 11 days, all treated lines had decreased to approximately 10^6 PFU/ml, but none of the four treated lines were inviable (Figure 2B).

Burst Time

Untreated phage populations increased rapidly after 18 minutes of incubation with exponentially-dividing hosts. After 24 hours, each of the three bisulfite-exposed populations showed a stable phage concentration for approximately 22-24 minutes, after which time each increased in density rapidly, until all exceeded 3500 PFU/ml after 28 minutes (Figure 3). After 48 hours of bisulfite exposure, each population remained stable for 26-32 minutes after exposure to hosts. Density began to increase slowly during that time, but did not increase sharply, indicating slower and more varied burst times in the phage population (Figure 3). These results indicate both a slowing and a diversifying of burst times following 24 and 48 hours of bisulfite treatment.

Sequence Data

PhiX174 populations were sequenced to an average depth of coverage of 65,000x. Untreated, 0.2M, and 0.4M bisulfite treated populations contained extensive and approximately equal levels of polymorphism (Figure 4). There were differences in the levels of polymorphism between the bases in treated an untreated samples, and between 0.2M and 0.4M treated populations (Figure 5). Treatment with 0.2M bisulfite significantly increased the frequency of 4 mutations compared to untreated populations (one-tailed T-test, $p<0.025$): A→T, G→T, T→A, and T→G, and 0.4M bisulfite treatment significantly increased the frequency an additional two: A→C and G→C.

E. coli Treatment

There were no differences in viability between untreated *E. coli* cultures and those exposed to 0.02M and 0.04M bisulfite. Qualitatively, when plated in TK top agar with and without bisulfite, but in the absence of phage, no differences in bacterial lawns were apparent between treated and untreated bacteria (Figure 6). Similarly, cell density (CFU/ml) of treated and untreated cultures did not significantly differ between treated and untreated populations (Figure 7).

In addition to having no discernable effect on viability, growth curves of treated and untreated *E. coli* indicated no significant effect of 0.02M and 0.04M bisulfite exposure on growth rate ($p>0.05$ for each timepoint). Treated and untreated cultures also reached the same density (10^6-10^7 CUF/ml) after six hours of growth, which remained constant for the duration of the experiment (Figure 8).

Electron Microscopy

Transmission electron microscopy revealed no discernable effects of bisulfite treatment on phiX174 capsid structure; intact capsids were evident in the control population and also in the 0.2M and 0.4M bisulfite treated populations (Figure 9), indicating no significant effects of bisulfite treatment on capsid integrity.

Discussion

ssDNA viruses, which mutate and evolve at rates comparable to RNA virus, present unique challenges for prevention and treatment. Mutagenesis has long been considered a potential treatment for infections of rapidly mutating and otherwise difficult-to-treat fast-evolving viruses. Mutagens are used in the treatment of some viral infections (Mangia et al., 2005), and experimentally, mutagenic treatment can have strong negative effects on viral fitness, implying that mutagenesis is a feasible treatment option against pathogenic viruses (Hayatsu and Miura, 1970; Pauly et al., 2015).

While viral control through mutagenesis is not a novel approach, the use of a cytosine specific mutagen to target and exacerbate a specific spontaneous mutation is. We have demonstrated that exposure to a cytosine specific mutagen can successfully drive populations of ssDNA bacteriophages to extinction. However, we cannot conclude this was achieved through lethal mutagenesis.

In two previous studies of mutagenesis in bacteriophages, one using 5-fluorouracil (5FU) against phiX174 and the other using nitrosoguanidine (NG) against dsDNA T7, fitness of treated populations improved (Domingo-Calap et al., 2012; Paff et al., 2014). In both cases, it was hypothesized that increasing mutation rate allowed for the phages to sample beneficial mutations, potentially one or more that conferred resistance to mutagenesis (Domingo-Calap et al., 2012). In both cases the mutagens used caused a wide range of mutations: 5FU indirectly encourages thymine to be replaced by any other base, and NG encourages transitions (Domingo-Calap et al., 2012; Ohnishi et al., 2008). By targeting a more narrow mutational target (just cytosine transitions), which are likely already at high frequency from spontaneous oxidative damage, we believe we have minimized the likelihood that treatment would lead the viruses to uncover beneficial mutations.

Recombination may also explain the fitness increases observed in these previous mutagen studies. In one previous study, T7 populations were more numerous than host cells during mutagenic treatment, allowing for coinfection and recombination, which may have countered the deleterious affects of mutagenesis by clearing mutations from subsequent generations (Paff et al., 2014). Our technique eliminated the potential for recombination and selection for beneficial variants in two ways. First, treatment was of virions independent of host cells, so mutations were able to accumulate in a neutral fashion, rather than constantly being subjected to replication and potential genomic recombination. Second, our plating technique maintained host populations far in excess of phage (approximately 10^5 CFU vs 10^1-10^2 PFU) minimizing the likelihood of coinfection and recombination during phage viability assays.

Our use of a cytosine-specific mutagen led to a extinction of the target phage populations. ssDNA experiences spontaneous cytosine deamination far more frequently than dsDNA (Frederico et al., 1990), and this mechanism may explain why ssDNA phages experience mutations rates closer to those of RNA viruses than dsDNA viruses. ssDNA phages, therefore, exist much closer to the lethal mutagenesis threshold than the larger, more stable dsDNA phages, and, through unstable, unpaired cytosine, provide an attractive target for deaminating agents. These two factors, combined with the very small and dense genomes of ssDNA viruses, make them appealing candidates for treatment via lethal mutagenesis of cytosine, in contrast to the larger, less dense, and slower-mutating genome of dsDNA viruses. The mammalian immune system exploits these properties to counter fast-evolving viruses with the APOBEC3 family of cytodine deaminases. These are part of the innate immune system and have been shown to induce hypermutation in retroviruses and RNA viruses (Fehrholz et al., 2012; Lee et al., 2008).

Mutagenesis has been used previously to drive viral populations to extinction, but despite these apparent successes, it is not clear that mutagenesis has ever actually been the mechanism through which a population is eliminated, either in vivo or in vitro. Recent work suggests that while mutagenesis may play a role in successful eradication of target viral populations, other factors are likely to contribute to the efficacy of mutagenic treatments (Bull et al., 2013; Graci and Cameron, 2008).

The most likely alternative explanation for the observation that viral fitness decreases during mutagenic treatment is that their hosts are also impacted, which has the secondary effect of slowing viral replication (Bull et al., 2013). If viral populations are treated while replicating in their hosts, the host is also exposed to the treatment, and the effects on the host can drive apparent fitness effects in the viruses. Past tests of mutagenesis in bacteriophage populations suffer from this possibility, as the viruses have been treated while replicating in host populations (Domingo-Calap et al., 2012; Paff et al., 2014; Pauly et al., 2015). Mutagenized hosts may suffer not only decreased fitness due to the accumulation of deleterious mutations, but also from the other affects of mutagens, which are often considerable (Domingo-Calap et al., 2012).

Our experimental design addressed this concern in two ways. First, by using sodium bisulfite, we minimized damage to the host cells. Bisulfite is mutagenic in both viruses and bacteria, but affects bacteriophages at lower concentrations than their hosts (Hayatsu and Miura, 1970). In preliminary assays, a significantly higher concentration was required to elicit a detectable effect on *E. coli* viability compared to phage viability (data not shown). Bacteria also have efficient mechanisms for the repair of deamination. For example, both the methyl-directed mismatch repair pathway and uracil and thymine N-glycosylase efficiently detect and repair C→T and C→U transitions in bacterial chromosomes (Krokan et al., 2002; Modrich, 1989; Visnes et al., 2009).

Second, during phage viability and growth assays, *E. coli* was not exposed to the same concentration of bisulfite as treated phage populations. By treating virions directly, independent of the host, and then evaluating viability or growth, we imposed a dilution of at least ten-fold from the mutagenic treatment to any growth assay involving host cells. At one-tenth the treatment concentration, neither viability nor growth rate of *E. coli* was effected (Figures 6-8). The very small differences in viability between treated and untreated T4 (Figure 2), which were enumerated on the same host as phiX174, further bolster the case that that any fitness changes observed in phiX174 populations were not the result of mutagenic treatment of their *E. coli* hosts.

It has been hypothesized that phage populations might not be strongly affected by mutagenic treatment because, when encapsidated, their densely packaged genomes may leave little room for mutagens to interact. Consequently, only the most loosely packaged outer regions of the genome could be affected, leading to mutational hot and cold spots, compromising the efficacy of mutagenesis. This was proposed as a possible explanation for an increase in fitness observed in T7 after treatment with hydroxylamine, although it does not appear to have completely excluded the mutagen from any genomic regions (Bull et al., 2013). T7 genomes are packaged near crystalline density within the capsid head (Earnshaw and Casjens, 1980), which could contribute to this problem. Conversely, phiX174 genomes are relatively unordered with the capsid (Benevides et al., 1991; Incardona et al., 1987), which should negate the difficulty of uniform exposure to the mutagen if virions are treated in the absence of hosts. Other ssDNA viral genomes are similarly unordered when encapsidated (Welsh et al., 1998; Wen et al., 1999), suggesting this may be a universal property of encapsidated ssDNA viruses. While recent work suggests ssDNA genomes may contain more secondary structures than previously thought (Muhire et al., 2014), though this does not necessarily mean the genomes are sufficiently dense to exclude chemical mutagens.

In many other tests of mutagenesis against viral populations, base analogs were used to increase mutation rate (for example (Domingo-Calap et al., 2012; Hayatsu and Miura, 1970; Paff et al., 2014; Pauly et al., 2015)). These operate during DNA synthesis: substitutions and mispairing may occur prior to genome replication, but they are “locked in” during DNA replication. By targeting cytosine deamination, a spontaneous chemical process that occurs independent of replication, we decoupled mutagenesis from DNA synthesis. This freed our analysis of the confounding factors of replication-associated mutation repair mechanisms, population growth, and the effects of mutagenic treatment on the viral host. In natural populations, these processes are operating, which has implications for the application of mutagenesis, but in vitro, they complicate efforts to quantify the fitness effects of mutagenesis.

Base analogs may also affect a wide range of processes in both the virus and host cell. For example, 5-fluorouracil, used in an earlier test of mutagenesis in phiX174, causes severe problems for the host cells exposed to the treatment. Detrimental affects beyond mutagenesis include interfering with metabolism, altering transcription, translation, and DNA synthesis, and thymine starvation, which itself leads to DNA breaks and error-prone DNA repair (Domingo-Calap et al., 2012). These effects, rather than mutagenesis, can contribute to loss of viral fitness, which can compromise efforts to demonstrate the efficacy of mutagenesis as an antiviral treatment.

An alternative explanation for loss of viability is that in addition to causing mutations, mutagenic treatment can cause DNA shearing. Hydroxylamine, another deaminating agent, shears DNA (Rhaese and Freese, 1968). DNA shearing would explain fitness costs associated with treatment, as fragmented phage genomes would be much less likely to effectively express the genes required for replication, especially with little or no coinfection, as was the case in these experiments. However, DNA shearing has not been observed with sodium bisulfite. Furthermore, bisulfite is commonly used to characterize methylation patterns in genomic DNA through deamination and sequencing, and DNA shearing is not considered a pitfall of these techniques (Darst et al., 2010; Patterson et al., 2011).

Finally, mutagens might interact with proteins and affect virus viability in ways that do not involve their disruption of genomic information. If bisulfite treatment disrupts protein structure, it could potentially affect the integrity of the viral capsids during treatment, preventing them from infecting host cells, and driving fitness down independent of mutagenesis. However, this is unlikely for several reasons. While there are obviously significant differences in capsid structure between phiX174 and T4, a general mechanism affected protein structure would be expected to affect both. Since T4 was relatively unaffected by treatment with 0.4M bisulfite, it is unlikely that this concentration unduly affects protein structure or stability. It is possible that bisulfite interacts with proteins in some specific way, and is able to react with a component of the phiX174 capsid but not a structural component of T4. However, bisulfite is only known to affect one kind of protein structure: it can inhibit the formation or promote the cleavage of disulfide bridges (Abtahi and Aminlari, 1997; Rom et al., 1992; Zhang and Sun, 2008). This is primarily exploited in food processing, where bisulfite is used to make starch in certain grains more bioavailable, as starch granules are often stored within a protective proteinaceous capsule stabilized by disulfide bonds. Bisulfite treatment frees the starch from this capsule (Choi et al., 2008). The phiX174 capsid does not contain any disulfide bridges, so there is no known mechanism by which bisulfite could affect the structure of the phiX174 capsid. Transmission electron microscopy confirmed the presence of intact capsids in populations of phages treated with bisulfite for two weeks, which indicates that capsid structure was unaffected (Figure 9).

Despite the evidence phage inactivation cannot be explained by a mechanism other than mutagenesis, and the significant difference in polymorphism levels between treated and untreated populations, our data do not indicate that mutagenesis alone can explain the loss of fitness in our treated phage populations. On average, about 0.080% reads were polymorphic in untreated populations. Populations treated with 0.2M and 0.4M bisulfite for two weeks contained 0.082% and 0.091% polymorphic reads, respectively. The phiX174 genome contains 5386 bases, so untreated populations contained an average of about 4.3 polymorphisms per genome, while treated populations contained just 4.4 (0.2M treated) and 4.9 (0.4M treated) polymorphisms per genome. Since the increase in polymorphism correlating with mutagenic treatment was less than one mutation per genome for both treatment levels, increased mutations cannot completely explain the loss of fitness in treated populations.

High-throughput next-generation sequencing was required to characterize underlying polymorphism within each population, but there are relatively high error rates associated with it, which are closely correlated with library preparation and sequence assembly techniques (Schirmer et al., 2015). Quantification and correction of these errors is difficult because they are often associated with read position (Cox et al., 2010; Schirmer et al., 2015), and in our populations, the average depth of coverage was approximately 65,000. To minimize the impact of these errors in our analysis, we eliminated from consideration any site that was not sequenced to a depth of coverage of at least 10,000, which should have removed the sites most influenced by sequencing error. However, our results show evidence for sequencing-related polymorphism, such as A/C and G/T misreads due to similar emission spectra, which have been identified as common problems associated with Illumina sequencing (Schirmer et al., 2015).

These confounding results could also have been influenced by sample preparation. This is one of the first studies to use Illumina to sequence ssDNA phiX174, in contrast to prior work that utilized a double-stranded form of the genome (Dickins and Nekrutenko, 2009). Control reads for Illumina are also carried out with phiX174, but with the commercially available, double-stranded RFII form. The use of ssDNA was necessitated by our treatment protocol, in which virions were exposed to the mutagen independent of the host; subsequent infection and purification of replicative, dsDNA phiX174 genomes would eliminate non-viable sequences from the survey. Consequently, normal library preparation methods could not be used. For example, to facilitate the capture of single-stranded phiX174 DNA, an RNA library prep kit was employed rather than a DNA kit during Illumina library preparation. Despite this change, a high percentage (~85%, personal communication Udi Zelzion) of reads mapped to the *E. coli* genome, indicating that efforts to bias sequencing towards ssDNA were not successful. Therefore, a significant proportion of our sequenced phiX174 may have been the result of free, double-stranded genomes which were not adequately removed. More stringent purification of virions followed by sequencing would eliminate this possibility.

Although mutagenic treatment of virions independent of hosts facilitates simplified analysis of the effects of such treatment, it can give an erroneously optimistic picture of the efficacy of mutagenesis as a treatment for viral infections. For example, cellular life has efficient mechanisms to recognize and repair cytosine deamination, such as uracil N-glycosylases (Krokan et al., 2002; Visnes et al., 2009), which could aid in viral survival when replicating in the presence of deaminating agents. These pathways may dictate the need for a higher dose of mutagen to achieve lethal mutagenesis, but given the relatively low concentration of bisulfite that phiX174 populations were able to tolerate, we would expect this treatment to be effective against replicating ssDNA viruses as well.

Sodium bisulfite is a more attractive mutagen than most for antiviral therapy for several other reasons. First, it is in widespread use and its safety is widely accepted. It has FDA GRAS certification, and though it has been documented to trigger allergic reactions in rare cases, it is not considered dangerous nor mutagenic to multicellular eukaryotes (Nair et al., 2003). Second, while other mutagens, such as base analogs, can affect a wide range of processes in exposed cells, bisulfite is narrowly limited in its activity to cytosine deamination.

ssDNA viruses present unique threats to global food security, and are rapidly emergent due to their fast rate of evolution. Like RNA viruses, they may be difficult targets for vaccination or more traditional anti-viral therapies. Mutagenesis may become a front-line weapon against viral infections in plant and animals, but its efficacy has never been conclusively documented. We have demonstrated that treatment with a cytosine-specific mutagen can eliminate ssDNA bacteriophage populations, and though we have eliminated many of the most likely alternatives, we cannot say that mutagenesis alone is responsible for the efficacy of the treatment.

Acknowledgements

We thank Lillian Nichols for collecting the T4 data, Natasia Jacko for assistance with burst time assays, Nicole Wagner and Udi Zelzion in the SEBS Genome Cooperative for Illumina sequencing and assembly, and Rajesh Patel in the Core Imaging Lab, Department of Pathology, Robert Wood Johnson Medical School, for the electron microscopy.

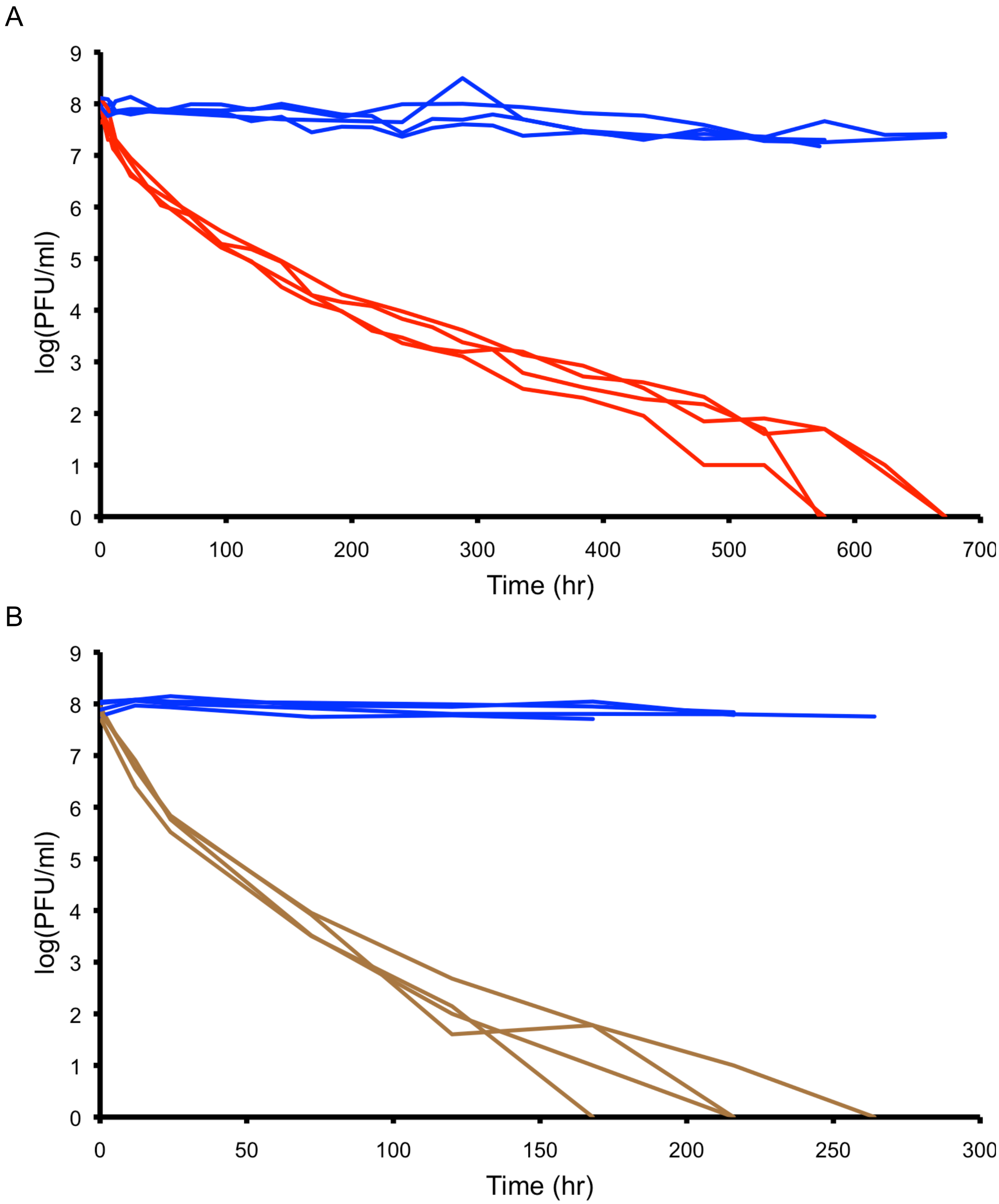


Figure 1. PhiX174 viability in A) 0M (blue) and 0.2M (red) sodium bisulfite and B) 0M (blue) and 0.4M (brown) sodium bisulfite, as measured in plaque forming units (PFU)/ml.

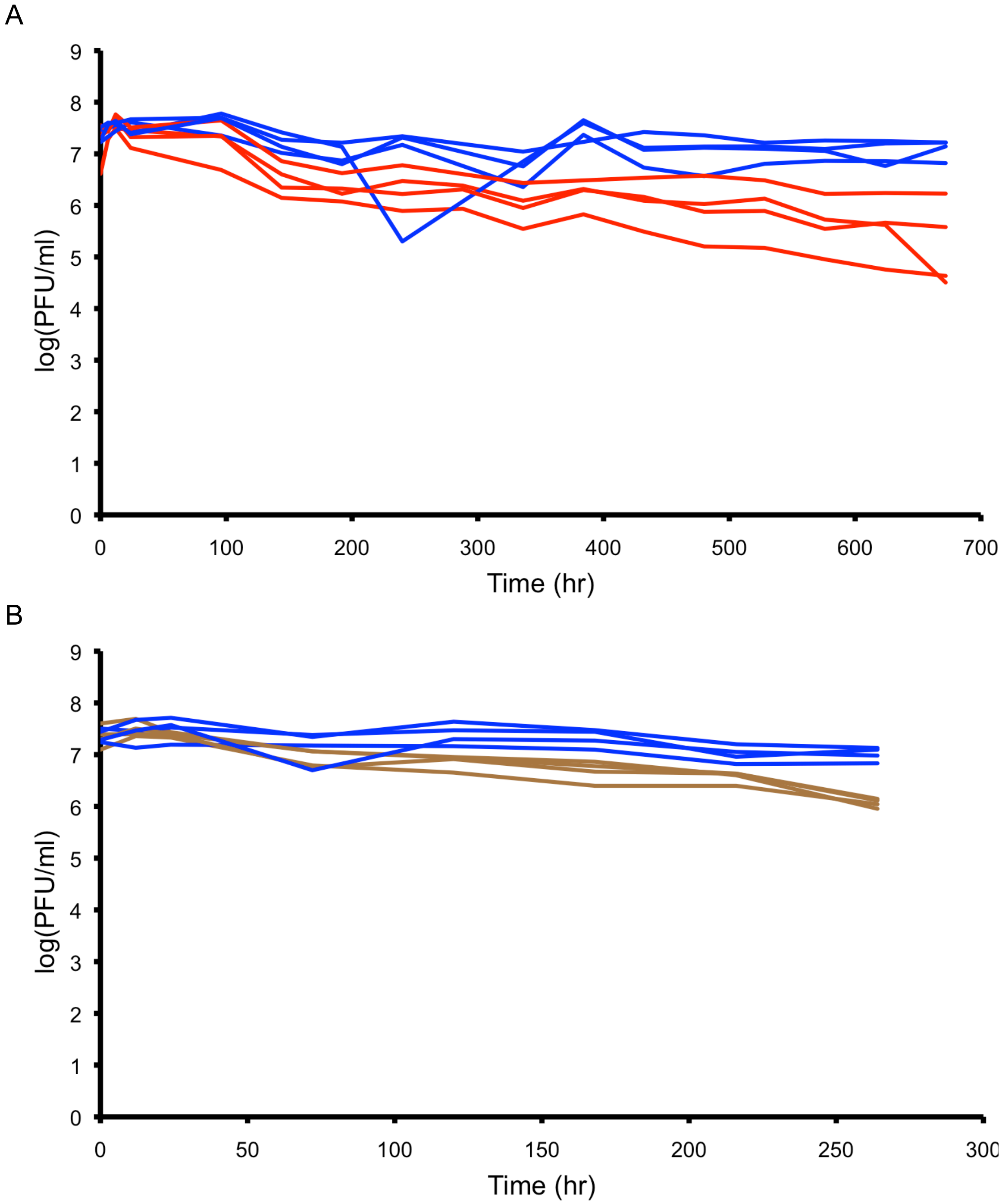


Figure 2. T4 viability in A) 0M (blue) and 0.2M (red) sodium bisulfite and B) 0M (blue) and 0.4M (brown) sodium bisulfite, as measured in plaque forming units (PFU)/ml.

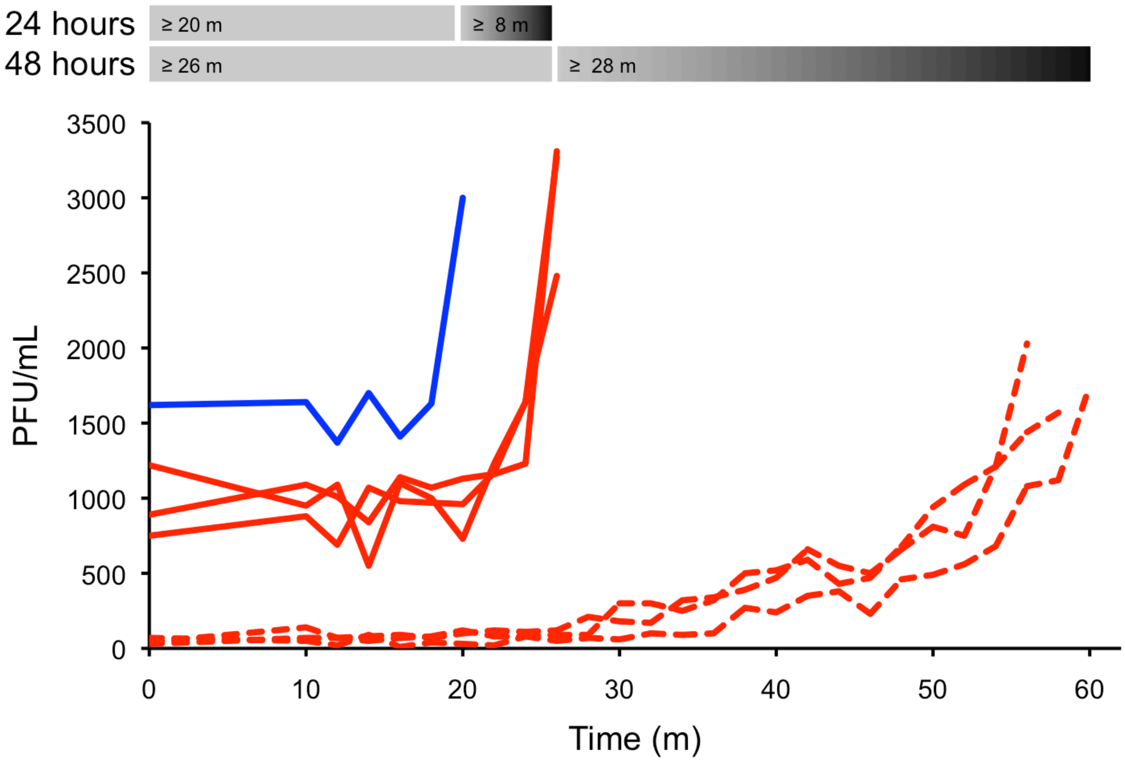


Figure 3. Sixty minute growth curves of phiX174 after 24 hours (solid red) and 48 hours (dashed red) of exposure to 0.2M sodium bisulfite. Bars across top indicate duration of lag (solid grey) and growth (gradient) phases for each trial; growth phase duration indicated as the time from the start of population growth to the point at which PFU/ml exceeded countability threshold. Blue line indicates growth curve for untreated phage population.

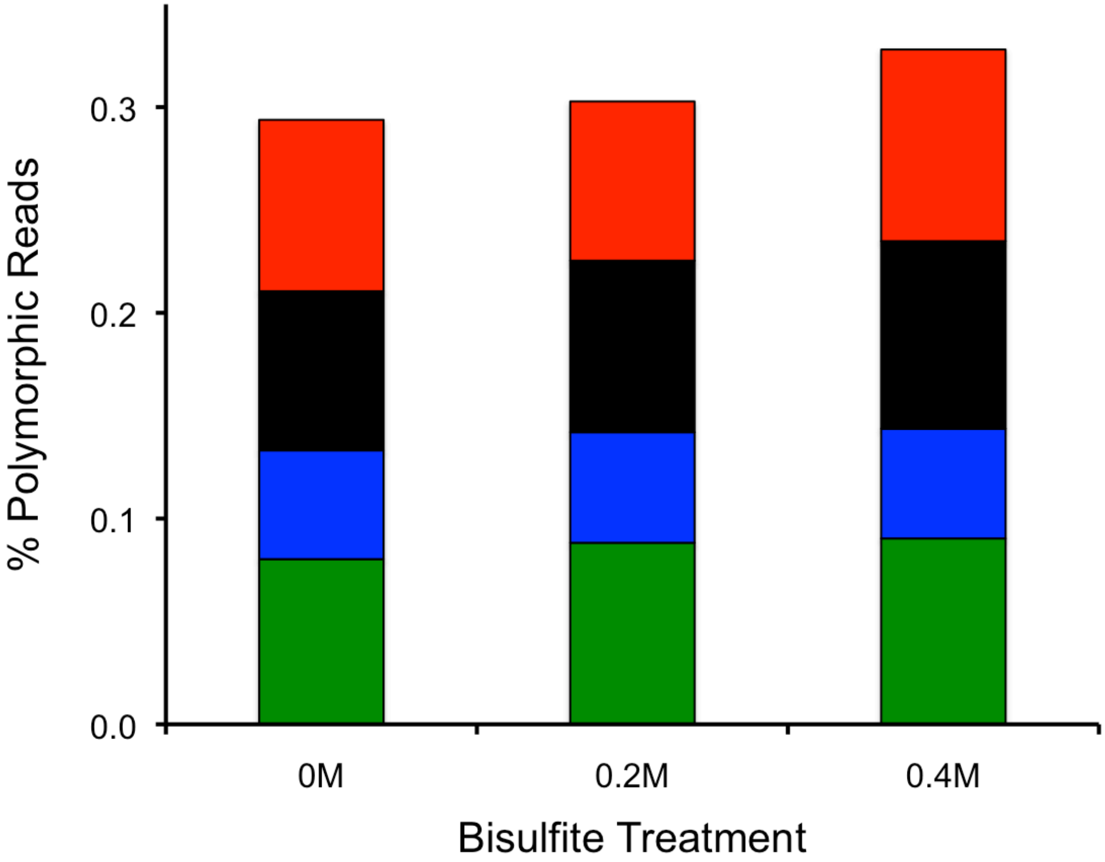


Figure 4. Percentage of polymorphic reads by consensus base according to Illumina sequencing in phiX174 populations after 14 days of exposure to 0M, 0.2M, or 0.4M sodium bisulfite. Color indicates consensus base, green is A, blue is C, black is G, red is T.

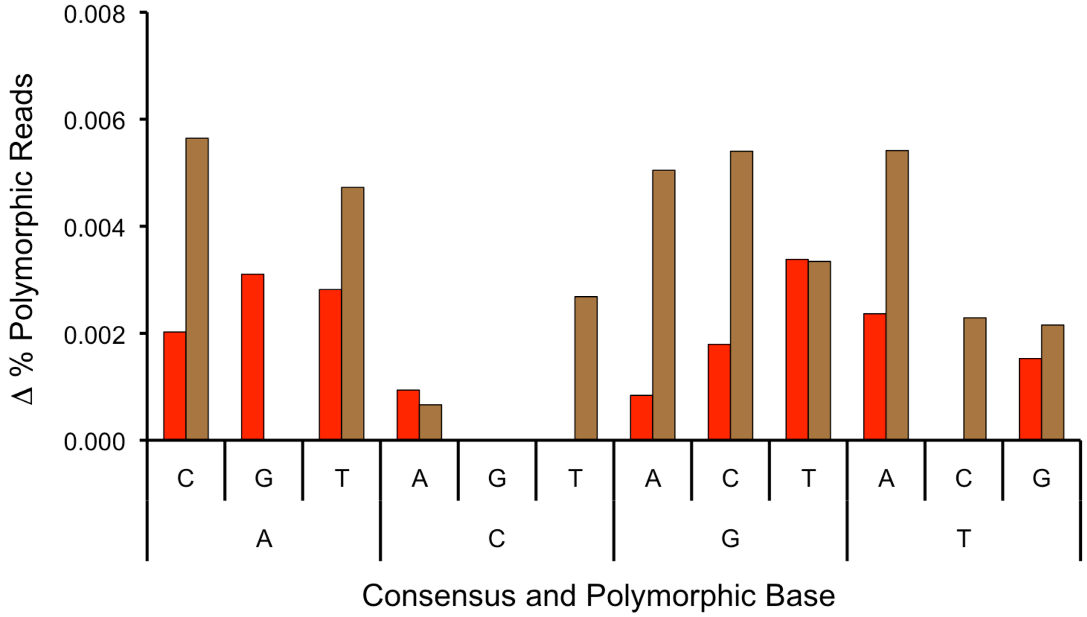


Figure 5. Change in the percentage of polymorphic reads compared to untreated populations for each consensus base (lower X axis labels) according to Illumina sequencing of phiX174 populations after 14 days of exposure to 0.2M (red), or 0.4M (brown) sodium bisulfite.

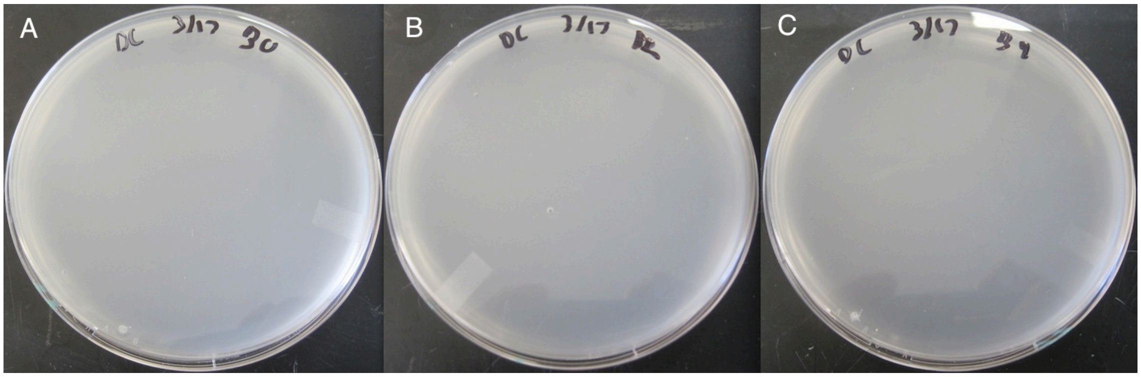


Figure 6. *E. coli* lawns in TK top agar on TK plates following exposure to A) 0M, B) 0.02M, and C) 0.04M sodium bisulfite and overnight incubation at 37°C.

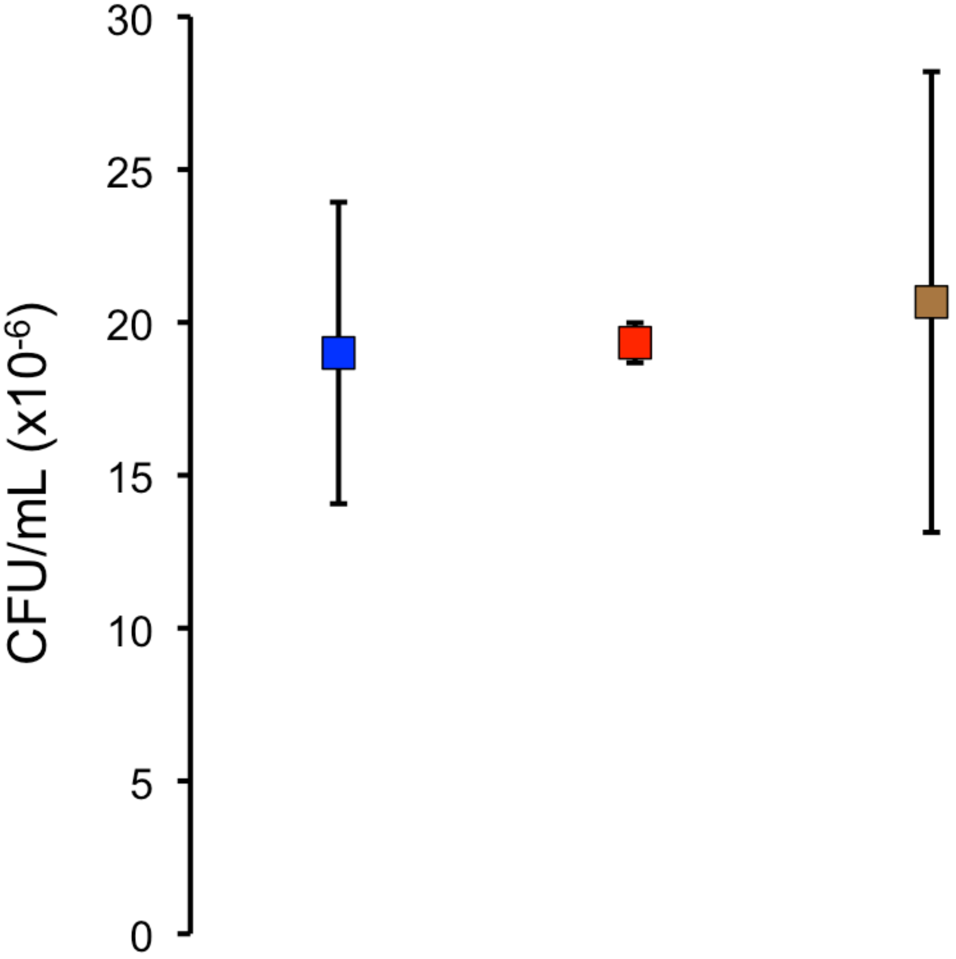


Figure 7. Colony forming units (CFU/ml) for *E. coli* treated overnight with 0M (blue), 0.02M (red), and 0.04M (brown) sodium bisulfite. Error bars indicate 95% confidence intervals.

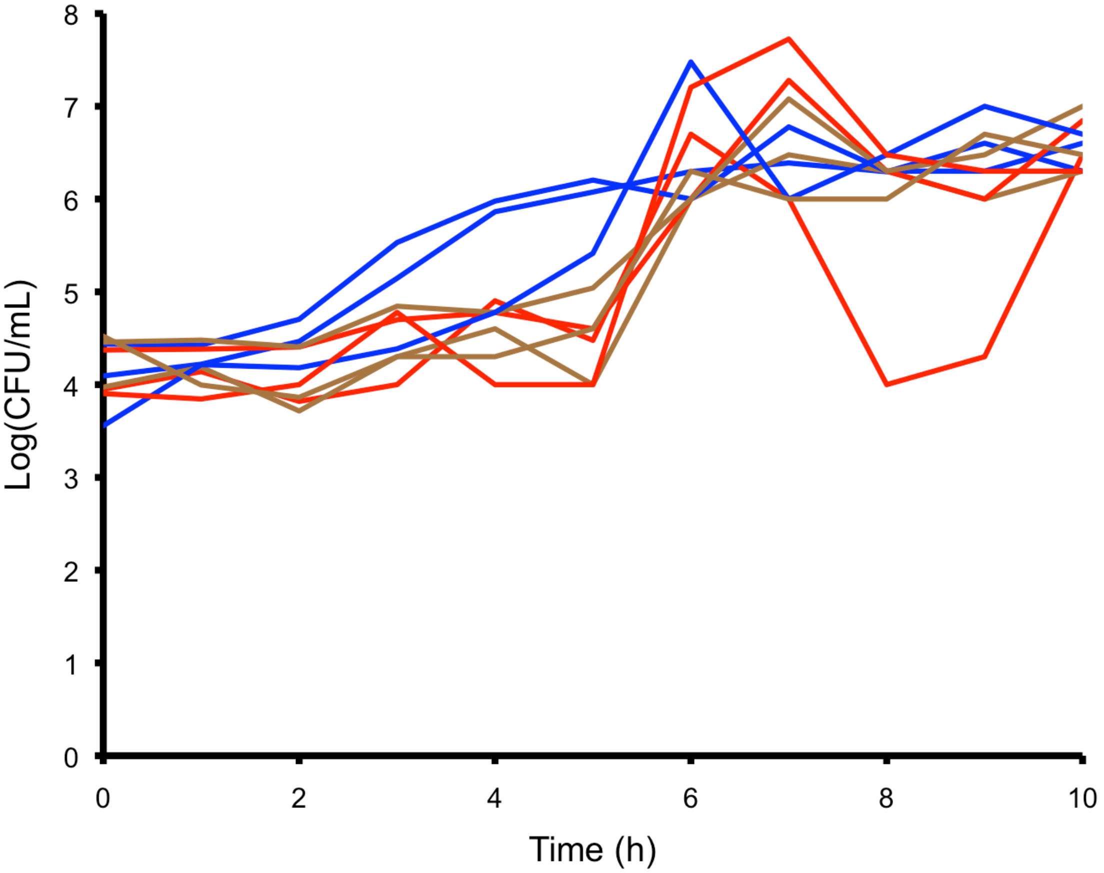


Figure 8. Ten hour growth curves for untreated (blue), 0.02M sodium bisulfite treated (red), and 0.04M sodium bisulfite treated (brown) *E. coli*.

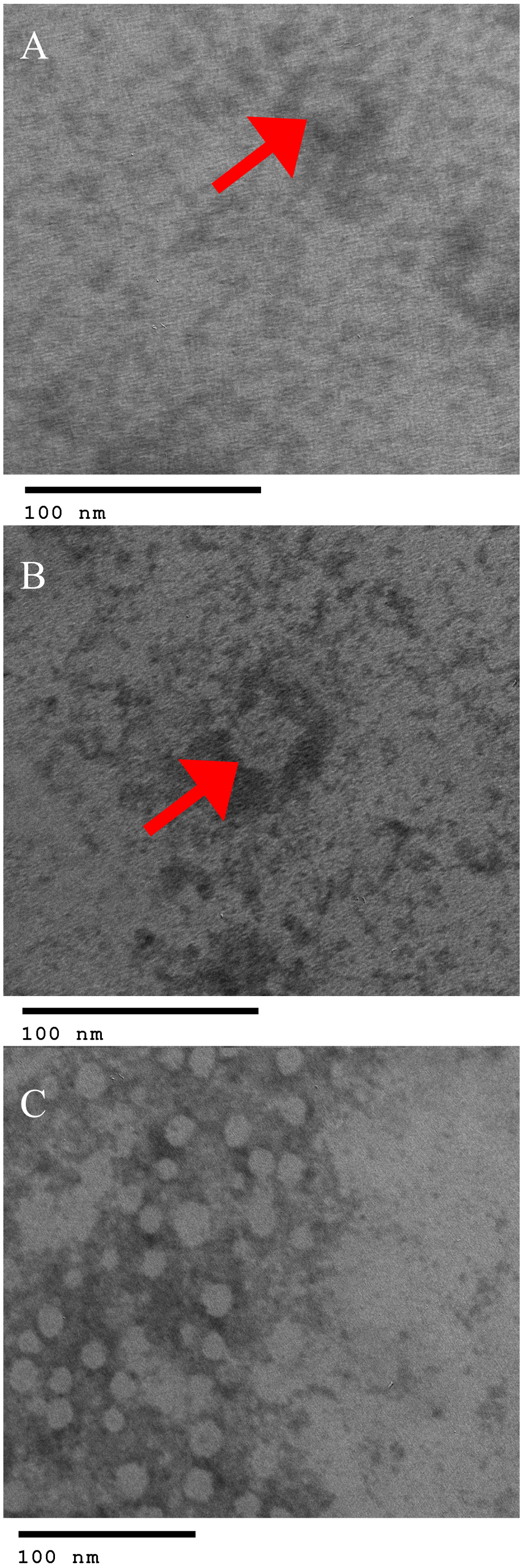


Figure 9. Transmission electron micrographs of A) untreated, B) 0.2M bisulfite treated, and C) 0.4M bisulfite treated phiX174.

References

Abtahi, S., Aminlari, M., 1997. Effect of Sodium Sulfite, Sodium Bisulfite, Cysteine, and pH on Protein Solubility and Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis of Soybean Milk Base. J. Agric. Food Chem. 45, 4768-4772.

Armstrong, D., Bishop, S., 2004. Does genetics or litter effect influence mortality in PMWS. Proceedings of the International Pig Veterinary Society Congress, 809.

Ball, C.L., Gilchrist, M.A., Coombs, D., 2007. Modeling Within-Host Evolution of HIV: Mutation, Competition and Strain Replacement. Bull. Math. Biol. 69, 2361-2385.

Benevides, J.M., Stow, P.L., Ilag, L.L., Incardona, N.L., Thomas, G.J., 1991. Differences in secondary structure between packaged and unpackaged single-stranded DNA of bacteriophage phi X174 determined by Raman spectroscopy: a model for phi X174 DNA packaging. Biochemistry (Mosc.) 30, 4855-4863.

Bodenheimer, J., Henry C. , Lindsay, K.L., Davis, G.L., Lewis, J.H., Thung, S.H., Seeff, L.B., 1997. Tolerance and Efficacy of Oral Ribavirin Treatment of Chronic Hepatitis C: A Multicenter Trial. Hepatology 26, 473-477.

Bull, J.J., Joyce, P., Gladstone, E., Molineux, I.J., 2013. Empirical Complexities in the Genetic Foundations of Lethal Mutagenesis. Genetics 195, 541-552.

Bull, J.J., Sanjuan, R., Wilke, C.O., 2007. Theory of Lethal Mutagenesis for Viruses. J. Virol. 81, 2930-2939.

Buonagurio, D.A., Nakada, S., Parvin, J.D., Krystal, M., Palese, P., Fitch, W.M., 1986. Evolution of Human Influenza A Viruses Over 50 Years: Rapid, Uniform Rate of Change in NS Gene. Science 232, 980-982.

Cardinale, D., DeRosa, K., Duffy, S., 2013. Base Composition and Translational Selection are Insufficient to Explain Codon Usage Bias in Plant Viruses. Viruses 5, 162-181.

Cardinale, D.J., Duffy, S., 2011. Single-stranded genomic architecture constrains optimal codon usage. Bacteriophage 1, 219-224.

Choi, S.J., Woo, H.D., Ko, S.H., Moon, T.W., 2008. Confocal Laser Scanning Microscopy to Investigate the Effect of Cooking and Sodium Bisulfite on In Vitro Digestibility of Waxy Sorghum Flour. Cereal Chemistry 85, 65-69.

Cox, M.P., Peterson, D.A., Biggs, P.J., 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11, 485.

Crotty, S., 2001. RNA virus error catastrophe: Direct molecular test by using ribavirin. Proceedings of the National Academy of Sciences 98, 6895-6900.

Cuevas, J.M., Pereira-Gómez, M., Sanjuán, R., 2011. Mutation rate of bacteriophage ΦX174 modified through changes in GATC sequence context. Infect. Genet. Evol. 11, 1820-1822.

Darst, R.P., Pardo, C.E., Ai, L., Brown, K.D., Kladde, M.P., 2010. Bisulfite Sequencing of DNA.

Dickins, B., Nekrutenko, A., 2009. High-Resolution Mapping of Evolutionary Trajectories in a Phage. Genome Biol. Evol. 1, 294-307.

Domingo-Calap, P., Pereira-Gomez, M., Sanjuan, R., 2012. Nucleoside Analogue Mutagenesis of a Single-Stranded DNA Virus: Evolution and Resistance. J. Virol. 86, 9640-9646.

Drake, J.W., 1991. A constant rate of spontaneous mutation in DNA-based microbes. Proc. Natl. Acad. Sci. U. S. A. 88, 7160-7164.

Duffy, S., Shackelton, L.A., Holmes, E.C., 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. 9, 267-276.

Earnshaw, W.C., Casjens, S.R., 1980. DNA packaging by the double-stranded DNA bacteriophages. Cell 21, 319-331.

Fane, B.A., Hayashi, M., 1991. Second-site suppressors of a cold-sensitive prohead accessory protein of bacteriophage phiX174. Genetics 128, 663-671.

Fehrholz, M., Kendl, S., Prifert, C., Weissbrich, B., Lemon, K., Rennick, L., Duprex, P.W., Rima, B.K., Koning, F.A., Holmes, R.K., Malim, M.H., Schneider-Schaulies, J.r., 2012. The innate antiviral factor APOBEC3G targets replication of measles, mumps and respiratory syncytial viruses. J. Gen. Virol. 93, 565-576.

Frederico, L.a., Kunkel, T.a., Shaw, B.R., 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry (Mosc.) 29, 2532-2537.

Graci, J.D., Cameron, C.E., 2006. Mechanisms of action of ribavirin against distinct viruses. Rev. Med. Virol. 16, 37-48.

Graci, J.D., Cameron, C.E., 2008. Therapeutically targeting RNA viruses via lethal mutagenesis. Future Virology 3, 553-566.

Hayatsu, H., Miura, A., 1970. The mutagenic action of sodium bisulfite. Biochem. Biophys. Res. Commun. 39, 156-160.

Incardona, N.L., Prescott, B., Sargent, D., Lamba, O.P., Thomas, G.J., 1987. Phage phi X174 probed by laser Raman spectroscopy: evidence for capsid-imposed constraint on DNA secondary structure. Biochemistry (Mosc.) 26, 1532-1538.

Krokan, H.E., Drabløs, F., Slupphaug, G., 2002. Uracil in DNA – occurrence, consequences and repair. Oncogene 21, 8935-8948.

Kunkel, T.A., 2004. DNA Replication Fidelity. J. Biol. Chem. 279, 16895-16898.

Lee, Y.N., Malim, M.H., Bieniasz, P.D., 2008. Hypermutation of an Ancient Human Retrovirus by APOBEC3G. J. Virol. 82, 8762-8770.

Legg, J.P., Fauquet, C.M., 2004. Cassava mosaic geminivirus in Africa. Plant Mol. Biol. 56, 585-599.

Majewski, J., 2003. Dependence of Mutational Asymmetry on Gene-Expression Levels in the Human Genome. The American Journal of Human Genetics 73, 688-692.

Mangia, A., Santoro, R., Minerva, N., Ricci, G.L., Carretta, V., Persico, M., Vinelli, F., Scotto, G., Bacca, D., Annese, M., Romano, M., Zechini, F., Sogari, F., Spirito, F., Andriulli, A., 2005. Peginterferon Alfa-2b and Ribavirin for 12 vs. 24 Weeks in HCV Genotype 2 or 3. N. Engl. J. Med. 325, 2609-2617.

Marti, T.M., Kunz, C., Fleck, O., 2002. DNA Mismatch Repair and Mutation Avoidance Pathways. J. Cell. Physiol. 191, 28-41.

McClelland, M., 1985. Selection Against dam Methylation Sites in the Genomes of DNA of Enterobacteriophages. J. Mol. Evol. 21, 317-322.

Modrich, P., 1989. Methyl-directed DNA Mismatch Correction. J. Biol. Chem. 264, 6597-6600.

Muhire, B.M., Golden, M., Murrell, B., Lefeuvre, P., Lett, J.-M., Gray, A., Poon, A.Y.F., Ngandu, N.K., Semegni, Y., Tanov, E.P., Monjane, A.r.L., Harkins, G.W., Varsani, A., Shepherd, D.N., Martina, D.P., 2014. Evidence of Pervasive Biologically Functional Secondary Structures within the Genomes of Eukaryotic Single-Stranded DNA Viruses. J. Virol. 88, 1972-1989.

Nair, B., Elmore, A., Panel, C.I.R.E., 2003. Final report on the safety assessment of sodium sulfite, potassium sulfite, ammonium sulfite, sodium bisulfite, ammonium bisulfite, sodium metabisulfite and potassium metabisulfite. Int. J. Toxicol. 22, 63-88.

Ohnishi, J., Mizoguchi, H., Takeno, S., Ikeda, M., 2008. Characterization of mutations induced by N-methyl-N,Ä≤-nitro-N-nitrosoguanidine in an industrial Corynebacterium glutamicum strain. Mutation Research/Genetic Toxicology and Environmental Mutagenesis 649, 239-244.

Paff, M.L., Stolte, S.P., Bull, J.J., 2014. Lethal Mutagenesis Failure May Augment Viral Adaptation. Mol. Biol. Evol. 31, 96-105.

Parrish, C.R., Kawaoka, Y., 2005. THE ORIGINS OF NEW PANDEMIC VIRUSES: The Acquisition of New Host Ranges by Canine Parvovirus and Influenza A Viruses. Annu. Rev. Microbiol. 59, 553-586.

Patterson, K., Molloy, L., Qu, W., Clark, S., 2011. DNA Methylation: Bisulphite Modification and Analysis. J. Vis. Exp.

Pauly, M.D., Lauring, A.S., Dermody, T.S., 2015. Effective Lethal Mutagenesis of Influenza Virus by Three Nucleoside Analogs. J. Virol. 89, 3584-3597.

Raney, J.L., Delongchamp, R.R., Valentine, C.R., 2004. Spontaneous mutant frequency and mutation spectrum for geneA of ?X174 grown inE. coli. Environ. Mol. Mutagen. 44, 119-127.

Rhaese, H.-J.r., Freese, E., 1968. Chemical analysis of DNA alterations: I. Base liberation and backbone breakage of DNA and oligodeoxyadenylic acid induced by hydrogen peroxide and hydroxylamine. Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis 155, 476-490.

Rom, D.L., Shull, J.M., Chandrashekar, A., Kirleis, A.W., 1992. Effects of Cooking and Treatment with Sodium Bisulfite on In Vitro Protein Digestibility and Microstructure of Sorghum Flour. Cereal Chemistry 69, 178-181.

Sanjuán, R., Nebot, M.R., Chirico, N., Mansky, L.M., Belshaw, R., 2010. Viral mutation rates. J. Virol. 84, 9733-9748.

Schimz, K.-L., 1980. The Effect of Sulfite on the Yeast Saccharomyces cerevisiae. Arch. Microbiol. 125, 89-95.

Schirmer, M., Ijaz, U.Z., D'Amore, R., Hall, N., Sloan, W.T., Quince, C., 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. 43, e37-e37.

Servey, J.T., Reamy, B.V., Hodge, J., 2007. Clinical Presentations of Parvovirus B19 Infection. Am. Fam. Physician 75, 373-376.

Steinhauer, D.A., Holland, J.J., 1986. Direct method for quantitation of extreme polymerase error frequencies at selected single base sites in viral RNA. J. Virol. 57, 219-228.

Visnes, T., Doseth, B., Pettersen, H.S., Hagen, L., Sousa, M.M.L., Akbari, M., Otterlei, M., Kavli, B., Slupphaug, G., Krokan, H.E., 2009. Uracil in DNA and its processing by different DNA glycosylases. Philosophical Transactions of the Royal Society B: Biological Sciences 364, 563-568.

Vivet-Boudou, V., Isel, C., El Safadi, Y., Smyth, R.P., Laumond, G., Moog, C., Paillart, J.-C., Marquet, R., 2015. Evaluation of Anti-HIV-1 Mutagenic Nucleoside Analogues. J. Biol. Chem. 290, 371-383.

Ward, C.D., Stokes, M.A., Flanegan, J.B., 1988. Direct measurement of the poliovirus RNA polymerase error frequency in vitro. J. Virol. 62, 558-562.

Ward, P., 2005. Oseltamivir (Tamiflu(R)) and its potential for use in the event of an influenza pandemic. J. Antimicrob. Chemother. 55, i5-i21.

Welsh, L.C., Marvin, D.A., Perham, R.N., 1998. Analysis of X-ray diffraction from fibres of Pf1 Inovirus (filamentous bacteriophage) shows that the DNA in the virion is not highly ordered. J. Mol. Biol. 284, 1265-1271.

Wen, Z.Q., Armstrong, A., Thomas Jr, G.J., 1999. Demonstration by ultraviolet resonance Raman spectroscopy of differences in DNA organization and interactions in filamentous viruses Pf1 and fd. Biochemistry (Mosc.) 38, 3148-3156.

Yule, T.D., Roth, M.B., Dreier, K., Johnson, A.F., Palmer-Densmore, M., Simmons, K., Fanton, R., 1997. Canine parvovirus vaccine elicits protection from the inflammatory and clinical consequences of the disease. Vaccine 15, 720-729.

Zhang, L., Sun, X.S., 2008. Effect of Sodium Bisulfite on Properties of Soybean Glycinin. J. Agric. Food Chem. 56, 11192-11197.