Differential selection and mutation shape codon usage of *Escherichia coli* ssDNA and dsDNA bacteriophages

Shivapriya Chithambaram

Supervisor : Dr. Xuhua Xia

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
University of Ottawa
In partial fulfillment of the requirements for a
Master's degree from the
Ottawa-Carleton Institute of Biology

Thèse soumise à la
Faculté des Etudes Supérieures et Postdoctorales
Université d'Ottawa
En vue de l'obtention de la maîtrise
L'Institut de Biologie d'Ottawa-Carleton

Abstract

Bacteriophages (hereafter referred as phages) can translate their mRNAs efficiently by maximizing the use of codons decoded by the most abundant tRNAs of their bacterial hosts. Translation efficiency directly influences phage fitness and evolution. Reengineered phages find application in controlling their host population in both health and industry. The objective of this thesis work is to examine the factors shaping codon choices of single stranded DNA (ssDNA) and double stranded DNA (dsDNA) *Escherichia coli* phages.

In chapter two, we employed two indices, r_{RSCU} (correlation in relative synonymous codon usage between phages and their hosts) and CAI (codon adaptation index) to measure codon adaptation in phages. None of the analyzed ssDNA phages encode tRNAs while some dsDNA phages encode their own tRNAs. Both r_{RSCU} and CAI are negatively correlated with number of tRNA genes encoded by these dsDNA phages. We observed significantly greater r_{RSCU} for dsDNA phages (without tRNAs) than ssDNA phages. In addition, we propose that ssDNA phages have evolved a novel codon adaptation strategy to overcome the disruptive effect of their high C→T mutation rates in codon adaptation with host.

In chapter three, we formulated an index φ to measure selection by host translation machinery and to present explicit linear and nonlinear models to characterize the effect of C→T mutation and host-tRNA-mediated selection on phage codon usage. The effect of selection (φ) on codon usage is detectable in most dsDNA and ssDNA phage species. C→T mutations also interfere with nonsynonymous substitutions at second codon positions, especially in ssDNA phages. Strand asymmetry along with the accompanying local variation in mutation bias can significantly affect codon adaptation in both dsDNA and ssDNA phages.

Résumé

Les Bactériophages (ci-après dénommés phages) peuvent traduire efficacement leurs ARNms en maximisant l'utilisation des codons décodés par les plus abondants ARNts de l'hôte bactérien. L'efficacité de la traduction influence directement l'évolution et la survie d'un phage. Les phages modifiés (par ingénierie) capable de contrôler la population de leur hôte, sont utilisés dans les domaines de la santé et de l'industrie. L'Objectif de cette thèse est d'examiner les facteurs qui influencent le choix de codons par l'ADN à brin simple (ADNss) et à brin double (ADNds) chez les phages d' *Escherichia coli*.

Au deuxième chapitre nous avons utilisé deux index afin de mesurer l'adaptation des codons chez les phages: le « r_{RSCU} » (la corrélation entre le phage et son hôte dans l'utilisation relative de codons synonymes) et le « CAI » (l'index d'adaptation de codon). Aucun des phages ADNss analysés codent des ARNts, tandis que certain phages ADNds codent leurs propres ARNts. Le r_{RSCU} et le CAI sont corrélés négativement avec le nombre de gènes ARNt codées par ces phages ADNds. Nous avons observé, de manière significative, des valeurs plus élevées du r_{RSCU} pour les phages ADNds (sans ARNts) que pour les phages ADNss. En outre, nous proposons l'évolution d'un nouveau codon comme stratégie d'adaptation chez les phages ADNss. Ceci, afin de surmonter les effets perturbateurs de leurs haut taux de mutation C→T lors de l'adaptation de codon avec leur hôte.

Dans le troisième chapitre, nous avons créé un index φ afin de mesurer le niveau de sélection effectué par la machinerie de traduction de l'hôte bactérien. Cet index a aussi été créé pour présenter d'explicites modèles linéaires et non-linéaires afin de caractériser l'effet des mutations C→T, et la sélection, contrôlée par les ARNts, sur l'utilisation des codons de phages. Les effets de la sélection (φ) sur l'utilisation de codons est détectable chez la majorité d'espèces de phages ADNds & ADNss.

En plus, les mutations C→T interfèrent avec les substitutions non-synonymes aux deuxièmes positions des codons, particulièrement chez les phages ADNss. L'asymétrie des brins et la variation du biais de mutation locale, peuvent affecter de manière significative l'adaptation des codons, chez les phages ADNss et ADNds.

Acknowledgements

I am extremely grateful to my supervisor Dr Xuhua Xia for providing me the opportunity to work in his lab as his master's student. I cannot thank Dr Xia enough for his immense support, constant encouragement and words of wisdom.

I would like to thank my advisory committee members, Dr Stéphane Aris-Brosou, Dr Ashkan Golshani and Dr Michel Dumontier for their useful inputs and guidance. In addition, I would also like to thank, Drs Stéphane Aris-Brosou, Ashkan Golshani and Nicolas Corradi for accepting to be my thesis examiners.

I take this opportunity to thank my past lab mates Anna and Manon for their huge support in helping me settle down in my first year. I must also thank Manon and Isabelle for generously accepting to do the French translation of my thesis abstract. I thank my lab mate Ramanandan for being very helpful while developing this thesis, for his useful suggestions and for reviewing my thesis multiple times. I also thank the other Xia lab members, Anwar and Akram for reviewing my thesis work.

I wish to thank all my friends for their support, in particular Megha and Nitya for suggesting me the idea of pursuing my higher studies in Canada and for being such dependable friends all these years. I thank my parents, my sister and the Almighty for giving me the strength and for surrounding me with positive energy, without their love and support the completion of this thesis would not have been possible. Finally, I would like to acknowledge the financial support from my dad, University of Ottawa and NSERC for their generous funding.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Abbreviations

A	Adenosine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
I	Inosine
Y	Pyrimidines (U/T and C)
R	Purines (A and G)
N	A, C, U/T and G
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
dsDNA	Double Stranded DNA
dsRNA	Double Stranded RNA
ssDNA	Single Stranded DNA
ssRNA	Single Stranded RNA
mRNA	messenger RNA
rRNA	ribosomal RNA
tRNA	transfer RNA
aaRS	aminoacyl tRNA synthetases
GTP	Guanosine triphosphate
GDP	Guanosine diphosphate
ATP	Adenosine triphosphate
HIV	Human immune deficiency virus
AA or aa	Amino acid
EF-Tu	Elongation factor Tu
CpG	Cytosine phosphate Guanine
CDS	Coding sequences
CF	Codon frequency
CP	Codon preferred
HEG	Highly expressed genes
LEG	Lowly expressed genes
NLM	Non-linear model
NCBI	National Center for Biotechnology Information
DAMBE	Data analysis in Molecular Biology and Evolution
ICTV	International Committee on Taxonomy of Viruses
gtRNAdb	Genomic tRNA database
RSCU	Relative Synonymous Codon Usage
CAI	Codon Adaptation Index
N_c	Effective number of codons

Amino acids abbreviations:

Ala	Alanine
Arg	Arginine
Asn	Asparagine
Asp	Aspartic acid
Cys	Cysteine
Gln	Glutamine
Glu	Glutamic acid
Gly	Glycine
His	Histidine
Ile	Isoleucine
Leu	Leucine
Lys	Lysine
Met	Methionine
Phe	Phenylalanine
Pro	Proline
Ser	Serine
Thr	Threonine
Trp	Tryptophan
Tyr	Tyrosine
Val	Valine

Ser_4	Serine four-fold group
Leu_4	Leucine four-fold group
Arg_4	Arginine four-fold group

1 Chapter One

Introduction

1.1 Protein synthesis

Protein synthesis is vital for survival and functioning of every organism. According to the central dogma of protein synthesis, the information stored in DNA is *translated* to proteins with the help of messenger RNA (mRNA). During translation, information stored in the mRNA is 'read' in groups of three nucleotides at a time. Such triplets of nucleotides located on the mRNA are known as 'codons'. Ribosomes function as the site of protein synthesis and catalyze the process of translation. Ribosomes are enzymatic complexes comprising two subunits. For instance, the bacterial ribosome is a 70S complex composed of a large 50S subunit and a small 30S subunit. Ribosomes consist of three sites for accommodating the incoming transfer RNAs (tRNA). They are, amino-acyl tRNA binding site (A-site), peptidyl site (P-site) and exit site (E-site). tRNA is the molecule responsible for serving the right amino acid to the appropriate codon present in the mRNA sequence. The process of adding the right amino acid to a tRNA is referred to as *charging* and the charging of tRNAs with their appropriate amino acids is performed by the enzyme tRNA-aminoacyl-synthetase.

The three main steps of protein translation are initiation, elongation and termination. Initiation of translation begins only when the tRNA corresponding to the start codon methionine, $tRNA^{fMet}$ is brought to the A-site of the ribosome where the start codon of the mRNA is positioned. The arrival of tRNA into the ribosome induces a conformational change in the ribosome, which is believed to go from an 'open' to 'closed' state causing the

16S RNA to come in contact with the minor groove of the codon-anticodon helix; this interaction monitors if correct codon-anticodon base pairing has occurred (Ogle et al. 2001; Ogle et al. 2002). If an incoming tRNA does not qualify to be the right match for the codon, then the tRNA is rejected as shown in Figure 1.1. Translational accuracy is said to be maintained through a kinetic proofreading process, which distinguishes between cognate and non-cognate tRNAs based on the differences in their free energy during binding process (Hopfield 1974; Ninio 1975).

A ternary complex comprising of amino acyl tRNA (for initiation it is $tRNA^{fMet}$), elongation factor Tu (EF-Tu) and guanosine triphosphate (GTP) is served to the ribosome. The correct codon-anticodon bindings result in the hydrolysis of GTP by EF-Tu. As a consequence of GTP hydrolysis EF-tu detaches from the aminoacyl tRNA and the tRNA proceeds to occupy the A-site of the ribosome. Upon the acceptance of the right cognate tRNA in the A-site, the ribosome translocates to read the next codon on the mRNA sequence. As a result, $tRNA^{fMet}$ is transferred to the P-site and this makes the next codon in the A-site available for the incoming tRNA to bind.

On the successful arrival of the next cognate tRNA to the A-site of the ribosome, a peptide bond is formed between the amino acid attached to the existing $tRNA^{fmet}$ at the P-site and the amino acid attached to the newly arrived tRNA at the A-site. At this stage, the elongating peptide chain continues to remain attached to the tRNA at the A-site while the ribosome moves to the read the next codon. As a consequence, the tRNA at the P-site (i.e., $tRNA^{fmet}$ which is uncharged now) moves to E-site and tRNA that was originally located at the A-site moves to the P-site along with the elongating peptide chain.

This movement is facilitated by two main steps. First, the cleavage of the hydrogen bond between the tRNA at the P-site and amino acid tethered to it and second is the

formation of a hydrogen bond between the dipeptide and the incoming tRNA that just translocated from the A-site to the P-site. This process continues until any one of the stop codons is encountered. Upon encountering a stop codon, a protein called release factor binds to the A-site of the ribosome instead of a tRNA and at this point the process of translation comes to a halt.

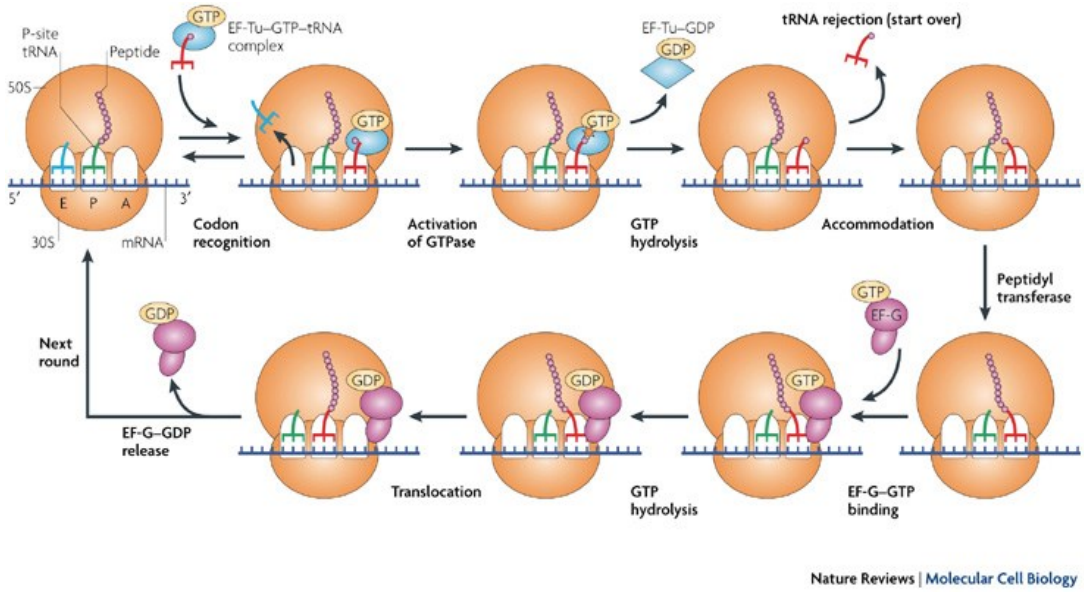

Figure 1.1 - An overview of steps involved in the translation of mRNA. Figure reproduced with permission from (Steitz 2008).

1.2 The standard genetic code

The genetic code is the template that determines which amino acid each codon should code for. The standard genetic code (Figure 1.2) contains 64 codons. Among these 64 codons, 61 encode amino acids and are known as sense codons. One of the sense codons coding for methionine (AUG) serves as the initiator of translation and is known as the start codon or initiation codon. The remaining three (UGA, UAG and UAA) codons are termed as stop or termination codons.

With the exception of methionine and tryptophan all other amino acids can be coded for by more than one codon. Hence the genetic code is said to be 'redundant' or '*degenerate*'. Codons coding for the same amino acid are grouped into codon families. Such codons are known as synonymous codons. Codon families are referred as single, two, three, four or six codon family based on the number of codons in each codon family. Apart from codons coding for Leu and Arg, all other codons belonging to a codon family differ only in the base at the third codon position. The two codon families have either A and G (purine ending) codons or C and Y (pyrimidine ending) codons at their third codon position. The isoleucine codon family is the only three codon family. This codon family has codons ending with all three bases except G at the third codon position. The four codon families contain codon variants with all four bases at their third codon position. Six codon families are formed as a result of combining two and four codon families.

Second letter

First letter	U	C	A	G	Third letter
U	UUU, UUC Phe; UUA, UUG Leu	UCU, UCC, UCA, UCG Ser	UAU, UAC Tyr; UAA Stop; UAG Stop	UGU, UGC Cys; UGA Stop; UGG Trp	U C A G
C	CUU, CUC, CUA, CUG Leu	CCU, CCC, CCA, CCG Pro	CAU, CAC His; CAA, CAG Gln	CGU, CGC, CGA, CGG Arg	U C A G
A	AUU, AUC, AUA Ile; AUG Met	ACU, ACC, ACA, ACG Thr	AAU, AAC Asn; AAA, AAG Lys	AGU, AGC Ser; AGA, AGG Arg	U C A G
G	GUU, GUC, GUA, GUG Val	GCU, GCC, GCA, GCG Ala	GAU, GAC Asp; GAA, GAG Glu	GGU, GGC, GGA, GGG Gly	U C A G

Figure 1.2 - The standard genetic code reproduced from a lecture slide of Dr Xuhua Xia.

Genetic codes with slight differences from the standard genetic code have been identified (Osawa et al. 1990). Exclusive genetic codes are available for species like viruses,

bacteria, fungi, and invertebrates. Genetic codes have also been documented for the cellular organelles like mitochondria and chloroplasts which differ marginally from the standard genetic code (Jukes, Osawa 1990). In total, 19 different types of genetic code have been deciphered so far (information retrieved from National Center for Biotechnology Information (NCBI 2013)).

1.3 Codon usage bias

Unlike the first two codon positions, changing the base at the third codon position can still yield the same amino acid in most cases. The usage of synonymous codons has been reported to be non-random among organisms. Codon usage bias is reported to be species-specific (Xia, Xie 2001), gene-specific (Gouy, Gautier 1982; Sharp et al. 1988) and in addition, differs even between different regions of a single gene (Duncan, Miller 1980). Selection and/or mutation are often invoked to explain the non-random codon usage in organisms. Ever since the empirical documentation of the positive correlation between tRNA abundance and codon usage (Barnes, Lindahl 2004), enormous progress has been made to understand the factors driving codon usage bias and codon-anticodon adaptation (Bulmer 1987; Xia 1998; Higgs, Ran 2008). For a fixed concentration of isoacceptor tRNAs in a species, codons forming the most efficient codon-anticodon pairs in terms of optimal binding energies are favored over other codons (Grosjean et al. 1978; Bennetzen, Hall 1982). Furthermore, highly expressed genes (HEGs) exhibit a greater extent of bias in codon usage than lowly expressed genes (LEGs) (Bennetzen, Hall 1982; Gouy, Gautier 1982; Sharp, Devine 1989). Thus, selection for increased translational efficiency (Coulondre et al. 1978; Ikemura 1981b; Robinson et al. 1984; Tuller et al. 2010) and translational accuracy (Grosjean et al. 1978; Robinson et al. 1984) appear to be shaping codon usage bias in

organisms. Despite increased attention to the role of selection in maintenance of codon usage bias, there have been limited studies examining the influence of mutation in shaping codon usage bias and the accompanying codon-anticodon adaptation. One such study that examined the effect of strand biased mutation in codon usage of vertebrate mitochondrial genomes revealed how mutation can facilitate codon-anticodon adaptation (Xia 2005).

Codon usage bias is not only genome specific but also tissue specific in the case of multicellular organisms (Plotkin, Robins, Levine 2004). For this reason, it becomes increasingly difficult to completely understand the factors affecting codon usage bias of an organism as the complexity of the organism increases. Hence, viruses and bacteria serve as the simplest and best model systems to better understand the causes of codon usage bias in individual organisms as well as to study effects of host on viral codon usage bias.

Viruses are intra-cellular obligate parasites. Thus, the destiny of viruses within their hosts is contingent upon how well they adapt to host imposed conditions and manage to escape host defense mechanisms. Viruses are constantly under selection pressure to match their genomic content according to host tRNA abundance (Sharp, Rogers, McConnell 1984) and mimic their host's genomic signatures (Greenbaum et al. 2008; Lobo et al. 2009). Viruses are classified into four major classes on the basis of their genetic material namely single-stranded DNA (ssDNA), double-stranded DNA (dsDNA), single-stranded RNA (ssRNA) and double-stranded RNA (dsRNA) viruses. In general, RNA viruses have very high mutation rates when compared to other viruses (Drake 1993). High mutation rates in RNA viruses have been attributed to the use of error-prone RNA dependent RNA polymerases (RdRp) which lack proof reading domains (Elena et al. 2000; Duffy, Shackelton, Holmes 2008). On the other hand, DNA viruses are known to have lower mutation rates since they use high fidelity DNA dependent DNA polymerases (DdDp) (Xia,

Yuen 2005; Duffy, Shackelton, Holmes 2008). Despite using high fidelity host DNA polymerases, the mutation rates of several ssDNA viruses have been reported to be similar to those of RNA viruses (Umemura et al. 2002; Shackelton et al. 2005; Shackelton, Holmes 2006; Duffy, Holmes 2008). Therefore, it is suspected that high mutation rates in ssDNA viruses may be an outcome of some underlying evolutionary process rather than simply polymerase error.

One of the chief contributors of spontaneous chemical modifications of DNA bases is the spontaneous deamination (Lindahl 1993). Among all other bases, cytosine is the most unstable in its unpaired state and is prone to spontaneous deamination rendering cytosine to be the hotspot for spontaneous mutations (Coulondre et al. 1978; Duncan, Miller 1980). If the C→T spontaneous mutations are not repaired prior to the next cycle of ssDNA replication, they leave their footprints in the genome by replacing the nucleotide C to T, and this modified base will then be used a template for next round of replication (Duncan, Miller 1980; Barnes, Lindahl 2004). Furthermore, studies have shown that C→T mutations mediated by spontaneous deamination is about 100 fold higher in ssDNA than in dsDNA (Frederico, Kunkel, Shaw 1990). Microorganisms having ssDNA genome have been proposed to have the highest per site mutation rate (Drake 1991). Such high spontaneous mutation rates in ssDNA phages have been postulated to affect phage adherence to host genomic nucleotide composition (Xia, Yuen 2005) and codon usage preferences (Cardinale, Duffy 2011). The genome size of ssDNA phages are generally very small. It should be noted that such small genome sizes in viruses indicate greater dependence on host for replication. In addition, none of the ssDNA phages encode their own tRNA genes. Thus, ssDNA phages solely depend on the host's translational system for their protein synthesis. Hence, selection pressure from host translational machinery will tend to force ssDNA phages to match their

codon usage to that of their hosts. On the other hand, high C→T mutation rates in ssDNA phages may interfere with phage codon adaptation. Our study aims to understand how ssDNA phages survive and replicate successfully within their hosts despite their high C→T mutation rates.

1.4 Methylation mediated spontaneous mutations and their effect on codon usage bias

Addition of a methyl (-CH3) group to DNA nucleotides is referred to as DNA methylation. DNA methylation is of the following two types. 1) Addition of methyl group to 5th position of pyrimidine ring of cytosine and 2) addition of methyl group at the 6th position of purine ring of adenine (Figure 1.3). Methylation of cytosine residue results in the formation of 5-methylcytosine and the methylation of adenine residue results in the formation of hypoxanthine.

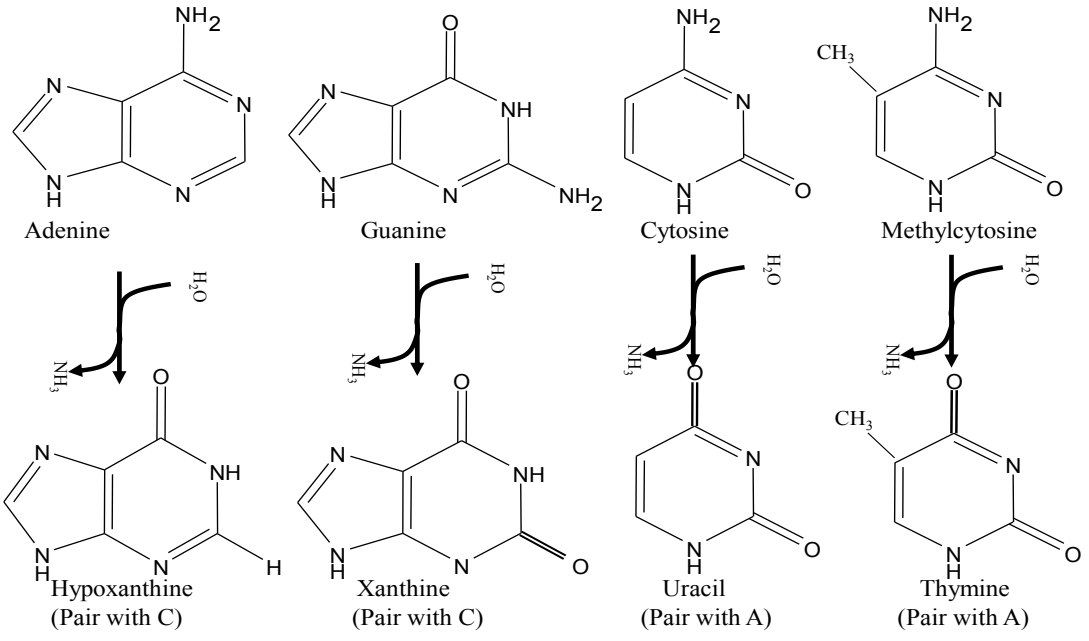

Figure 1.3 - Spontaneous deamination of nucleotides, figure reproduced from a lecture slide of Dr Xuhua Xia**.**

Spontaneous mutation rates are higher for 5-methylcytosine than cytosine, causing them to become hotspots for mutation (Lindahl 1993). Spontaneous deamination of cytosine results in thymine (if methylated) or uracil (in the absence of methylation). While cytosine to uracil deamination can be corrected by repairing enzyme uracil-DNA glycosylase, an equally efficient repair mechanism appears to be unavailable for correcting C→T deaminations (Gabbara, Wyszynski, Bhagwat 1994). Unchecked spontaneous deamination of methylated cytosine residues in the CpG dinucleotides leads to a steady decline in CpG frequencies and corresponding increase in TpG frequencies. DNA in single stranded state is claimed to be more prone to C→T mutations than double stranded DNA (dsDNA) (Frederico, Kunkel, Shaw 1990). Empirical evidence from bacteria revealed that the non-transcribed strand which is left single-stranded for a longer period is more prone to C→T mutations (Beletskii, Bhagwat 1996). Similarly, during replication of vertebrate mitochondrial genome, due to spontaneous deamination mutations, the L and H-strands are selectively enriched with A, C and G, T nucleotides respectively. Investigation of codon usage of protein coding genes from vertebrate mitochondrial genome revealed overrepresentation of A and C-ending codons in L-strand and G, T-ending codons in H-strand (Xia 2005). Nucleotide biases resulting from DNA methylation have been amply documented in viruses (Karlin, Doerfler, Cardon 1994; Shackelton, Parrish, Holmes 2006; Hoelzer, Shackelton, Parrish 2008). Spontaneous mutation mediated by methylation has been invoked to explain codon usage bias in several viruses (Woo et al. 2007; Cardinale, Duffy 2011). Such studies demonstrate the role of methylation mediated mutations in shaping genomic biases of viruses, thereby adding another dimension to the understanding of evolution of viruses. In the present study, we are particularly interested in studying the effect of spontaneous C→T mutations on codon adaptation of ssDNA and dsDNA phages.

1.5 Strand Asymmetry as a contributor to codon usage bias

According to Chargaff's parity rule 2 (Chargaff 1968), the number of A = T and C = G within each DNA strand. Deviations from this expectation occur due to biased effects of mutation or selection on the leading and lagging strand during DNA replication (Sueoka 1995); this is known as strand specific bias or strand asymmetry. One of the earliest works in strand asymmetry reported excess of G compared with C in leading strand and excess of C compared with G in lagging strand in several species (Lobry 1996). Differences in the replication mechanisms of the leading and the lagging strand (Marians, 1992) and the difference in DNA repair systems of the two strands (Francino et al. 1996) have been invoked to explain the asymmetry between the strands in bacterial species. Numerous studies have reported strand asymmetry in bacterial species (Francino, Ochman 1997; Karlin 1999; Lobry, Sueoka 2002; Xia 2012a) and dsDNA viruses (Mrazek, Karlin 1998; Kano-Sueoka, Lobry, Sueoka 1999).

GC skew and *AT skew* are two indices proposed to identify the presence of strand asymmetry (Lobry 1996). They are calculated using the following equations.

$$AT\ skew = \frac{A-T}{A+T} \tag{1.1}$$

$$GC\ skew = \frac{G-C}{G+C} \tag{1.2}$$

where *A*, *T*, *G* and *C* refer to the frequencies of the nucleotides *A*, *T*, *G* and *C* respectively.

Similarly, transcription and replication coupled strand bias has been reported to shape codon usage in phages (Kano-Sueoka, Lobry, Sueoka 1999). Strand asymmetry caused by differential mutation pressures during replication of leading and lagging strand can result in

biased codon usage in some dsDNA viruses (Mrazek, Karlin 1998). Such studies corroborate strand asymmetry as a significant contributor to codon usage bias.

1.6 Indices for measuring codon usage bias

1.6.1 Relative synonymous codon usage

Relative synonymous codon usage (RSCU) is a normalized index of codon usage bias. It measures the degree of usage of a particular codon in comparison with its alternate synonymous codons. The equation for computing RSCU was developed by Sharp and his co-workers (Sharp, Tuohy, Mosurski 1986) and it is presented below

$$RSCU_{ij} = \frac{X_{ij}}{\frac{1}{n_i}\Sigma_{j=1}^{n_i} X_{ij}} \tag{1.3}$$

In the above equation, X_{ij} denotes the observed frequency of the codon 'j' of the amino acid 'i', having 'n_i' synonymous codons that can code for amino acid "i". The maximum value that RSCU can take is the maximum number of codons in a codon family. A codon is said to be overused if it has a RSCU value greater than one and underused if it has value less than one. An RSCU value of zero would mean that the codon is not used at all.

1.6.2 Codon adaptation index

Codon adaptation index (CAI) was first proposed by Sharp *et al* to measure the codon usage bias of a gene (Sharp, Li 1987). Unlike RSCU which is a codon specific index, CAI is a gene-specific index of codon usage bias. The gene whose codon usage bias is to be measured is compared with a reference set (set of genes that are known to be highly expressed genes of the organism). So, the frequencies of codons in the gene whose CAI is to be measured and the frequencies of the codons in the reference set are pre-requisites for computing weight factor used in CAI calculation.

The initial implementation of CAI has few problems. They are 1) not handling the two and four-fold codon families separately and considering them jointly as a six-fold codon family, despite the fact that the two codon families are subject to different tRNA mediated selection pressures; 2) single codon families always have a CAI of 1 irrespective of bias, hence they must be excluded in CAI computation but previous software implementations unfortunately do not do so; 3) observed codon frequencies from gene sequences need to be directly used as input in the place of absolute numbers in CAI computation. The above mentioned problems have been taken care of in the latest version of CAI (Xia 2007) and it has been incorporated in the software DAMBE (Xia 2013b). All CAI analyses employed in this study were computed using DAMBE version 5.3.70 (Xia 2013b).

1.6.3 Effective number of codons

Effective number of codons (N_c) is a codon usage index developed by Frank Wright (Wright 1990). It measures the extent of deviation from even usage of synonymous codons. N_c values can range from a minimum value of 20 (extreme bias) when only a specific codon in each synonymous codon family has been used to code for an amino acid and up to a maximum value of 61 (when there is absolutely no bias) when all codons occur at equal frequencies. N_c is a gene-specific index of codon usage bias like CAI. The primary difference between the two indices (i.e., CAI vs N_c) being the fact that CAI computation requires a reference set while N_c does not.

The initial implementation N_c by Wright has a number of shortcomings. They are: 1) it does not handle the two and four-fold codon families distinctly, instead it treats them together as a six-fold codon family, despite the fact that the two codon families are subject to different selection pressures; 2) it has been developed based on the standard genetic code and

hence cannot be used to measure codon bias in the cases where the other available genetic code have to be used; 3) due to the absence of weights to distinguish between the different types of codon families has given the two codon families undue advantage in codon usage bias scores. The above mentioned deficiencies have been addressed in the enhanced version of N_c (Sun, Yang, X. 2013) which has been recently implemented in DAMBE (Xia 2013b).

1.7 Phage Biology and Classification

International Committee on Taxonomy of Viruses (ICTV) (Fenner 1971) is the international body dealing with the classification of viruses. Depending on whether phages have DNA or RNA as their genetic material, they are classified into DNA and RNA phages respectively. In addition, based on their genome architecture (strandedness), phages are classified into single stranded and double stranded phages. On the basis of nucleic acid type and strandedness, phages can be categorized into ssDNA, dsDNA, single stranded RNA (ssRNA) or double stranded RNA (dsRNA) viruses. In this work, we are particularly interested in studying codon usage patterns of ssDNA and dsDNA phages.

ssDNA and dsDNA phages can be further classified based on their morphologies and replication mechanisms in the following manner respectively. ssDNA phages are of two types, isometric and filamentous. dsDNA phages can either be lytic or lysogenic. dsDNA phages constitute the majority of the sequenced phages. More than 90% of the known dsDNA phages fall under the group of tailed phages (Ackermann 2007). There are three main families of tailed phages which belong to the order *Caudovirales.* They are *Myoviridae* (contractile tails), *Podoviridae* (short tails) and *Siphoviridae* (phages with long noncontractile tails).

The genome sizes of ssDNA phages are very small relative to the large dsDNA phages. Based on the sequence data collected from NCBI (National Centre for Biotechnology information), the size of dsDNA phages ranges from to 11 KB to 300 KB (data not shown). In contrast the size of the largest ssDNA genome is only between 4 to 10KB (data not shown). The large genome size of dsDNA phages allows them to carry some of their genes including some tRNA genes. While the genome size of ssDNA phages are quite low to accommodate genes which makes them almost entirely dependent on their host's machinery for replication.

1.8 Significance of the study

Phages are potential therapeutic agents that can be applied in controlling life-threatening bacterial infections in both plants and animals. Phage therapy refers to the application of phages to destroy pathogenic bacteria. The advent of multidrug resistance to several bacterial pathogens has resulted in considering phage therapy as an alternative to conventional antibiotics (Matsuzaki et al. 2005; Burrowes et al. 2011). This is due to the fact that phage therapy offers an effective alternative to anti-bacterials because phages are specific and cheaper to produce. Despite these benefits, phage therapy has not yet gained sufficient recognition in most western countries primarily due to the lack of clear understanding of phage biology. However, in the recent past many researchers have focused their attention to investigate the phages and understand a broader perspective of their evolution. The effect of codon optimization in viral genes i.e., the viral gene synonymous codon usage bias has been re-engineered in alignment to their host's codon usage bias to improve phage translation efficiency has been the area of major interest. Codon usage bias directly influences the gene expression level and helps in the synthesis of efficient phage

genomes that can escape host restriction enzymes (Skiena, 2001). Hence, codon adaptation studies facilitate better understanding of phage translation efficiency and thus helps re-design more efficient phages. Knowledge of phage translation efficiency is crucial for amplifying production of therapeutically and industrially important phages.

1.9 Research motivation

Although there is extensive published literature on codon biases of phages, studies relating phage lifestyle and genome organization (single or double stranded DNA) with the extent of codon adaptation with their hosts are rather limited. Our work aims to study the factors shaping codon usage choices of ssDNA and dsDNA phages. Furthermore, previous studies on phage codon adaptation are insufficient in two ways. First, they did not properly characterize the selection imposed by host translation machinery. Second, they do not clearly separate the effect of mutation and selection. The availability of complete genome sequences of a large number of phages has facilitated codon usage bias studies to be carried out in a large range of phages. This is the first work that formulated an index to measure selection by host translation machinery and to present explicit linear and nonlinear models to characterize the effect of C→T mutation and host-tRNA-mediated selection on codon usage.

1.10 Overview of subsequent chapters

This chapter introduces the process of protein synthesis, codon usage bias and includes literature review of the factors shaping codon usage bias.

In chapter 2, we compared the degree of codon adaptation between dsDNA phages and ssDNA phages with the help of codon usage indices RSCU and CAI. Here, we report our findings that dsDNA phages show better codon usage adaptation with their hosts than ssDNA phages. In contrast to some dsDNA phages none of the ssDNA phages harbor tRNA

genes in their genomes. We also assessed the contribution of phage encoded tRNAs in shaping phage codon adaptation. Finally, we propose a novel codon adaptation pattern exhibited by in ssDNA phages.

In chapter 3, we developed linear and non-linear models to elucidate the relative effect of mutation and selection on the codon usage of *Escherichia coli* phages. Here we show that, much of the variation in codon adaptation among dsDNA phages can be attributed to lineage effects, with some phage lineages having uniformly strong codon adaptation and some other lineages having uniformly weak codon adaptation (which may indicate recent host switching from other hosts to *E. coli*). Strand asymmetry with the associated local variation in mutation bias (e.g., U-biased in half of the genome and C-biased in the other half) can significantly interfere with codon adaptation in both dsDNA and ssDNA phages.

2 Chapter Two

Differential codon adaptation between dsDNA and ssDNA phages in *Escherichia coli*

2.1 Abstract

Because phages use their host translation machinery, their codon usage should evolve towards that of highly expressed host genes. We used two indices to measure codon adaptation of phages to their host, r_{RSCU} (the correlation in RSCU between phages and their host) and CAI computed with highly expressed host genes as the reference set. These indices used for this purpose are appropriate only when hosts exhibit little mutation bias, so only phages parasitizing *E. coli* were included in the analysis. For double-stranded (dsDNA) phages, both r_{RSCU} and CAI decrease with increasing number of tRNA genes encoded by the phage genome. r_{RSCU} is greater for dsDNA phages than for ssDNA phages, and the low r_{RSCU} values are mainly due to poor concordance in RSCU values for Y-ending codons between ssDNA phages and the *E. coli* host, consistent with the predicted effect of C→T mutation bias in the ssDNA phages. Strong C→T mutation bias would improve codon adaptation in codon families (e.g., Gly) where U-ending codons are favored over C-ending codons ("U-friendly" codon families) by highly expressed host genes, but decrease codon adaptation in other codon families where highly expressed host genes favor C-ending codons against U-ending codons ("U-hostile" codon families). It is remarkable that ssDNA phages with increasing C→T mutation bias also increased the usage of codons in the "U-friendly" codon families, thereby achieving CAI values almost as large as those of dsDNA phages. This represents a new type of codon adaptation.

2.2 Contributions

The data, results and interpretations in this chapter were submitted to *Molecular Biology and Evolution* journal. Shivapriya Chithambaram (SC) is the first author, Ramanandan Prabhakaran (RP) is the co-author and Dr Xuhua Xia (XX) is the corresponding author. This work was the result of a collaborative project between me and members of the Xia lab: RP and XX. This project was conceptualized by XX and provided supervisorial guidance. The development of the hypotheses, data analyses and interpretations resulted from discussions among SC, RP and XX.

2.3 Introduction

Efficient production of proteins is essential for survival and reproduction and strongly affects the fitness of a genotype, especially in unicellular organisms and viruses where rapid replication is essential for propagating the genotype into future generations. Efficient translation depends on the efficiency of the three sub-processes of translation, i.e., initiation, elongation and termination. Codon-anticodon adaptation directly impacts elongation efficiency. Ever since the empirical documentation of the correlation between codon usage and tRNA abundance (Ikemura 1981a), codon-anticodon adaptation has been well documented in bacterial and fungal genomes (Ikemura 1981a; Gouy, Gautier 1982; Ikemura 1992; Xia 1998) as well as in mitochondrial genomes in vertebrates (Xia 2005; Xia et al. 2007) and fungi (Carullo, Xia 2008; Xia 2008). In short, differential tRNA availability almost invariably leads to biased codon usage, with most frequently used codons corresponding to the most abundant tRNA species. Optimizing codon usage according to host codon usage has been shown to increase the production of viral proteins (Haas, Park, Seed 1996; Ngumbela et al. 2008) or transgenic genes (Hernan et al. 1992; Kleber-Janke,

Becker 2000; Koresawa et al. 2000). Studies on codon-anticodon adaptation have progressed in theoretical elaboration (Bulmer 1987; Bulmer 1991; Xia 1998; Higgs, Ran 2008; Jia, Higgs 2008; Xia 2008; Palidwor, Perkins, Xia 2010), in critical tests of alternative theoretical predictions (Xia 1996; Xia 2005; Carullo, Xia 2008; van Weringh et al. 2011) and in formulation and implementation of codon bias indices such as relative synonymous codon usage (RSCU, Sharp, Tuohy, Mosurski 1986), effective number of codons (Nc, Wright 1990; Sun, Yang, X. 2013) and codon adaptation index (CAI, Sharp, Li 1987; Xia 2007). Although a recent study has questioned the relationship between codon usage and protein production (Kudla et al. 2009), its conclusion has been found to be unwarranted (Tuller et al. 2010).

Bacteriophage (phage) needs to have efficient translation in order to survive among alternative phage genotypes. Because phages depend mainly on the translation machinery of their host for protein translation, their codon adaptation is shaped by mutation and selection of the host tRNA pool (Grosjean et al. 1978; Gouy 1987; Kunisawa, Kanaya, Kutter 1998; Sahu et al. 2005; Carbone 2008; Lucks et al. 2008). While some studies have suggested that extrinsic factors such as temperature (Sau, Deb 2009) and host diversity (Sau et al. 2007) may also affect phage codon usage, such factors should act indirectly through mutation and selection.

In order to study factors contributing to phage codon adaptation, we first use two codon usage indices, r_{RSCU} (correlation of RSCU values between the host and the phage) and CAI, to measure phage codon adaptation. As explained below, these indices are appropriate measures of phage codon adaptation when the host exhibits little nucleotide bias indicating little mutation bias. We then derive testable predictions on factors that contribute to phage codon adaptation.

2.3.1 Two codon usage indices to measure phage codon adaptation

Assuming that the codon usage of highly expressed host genes are well adapted to their own translation machinery, we expect the phage genes to evolve a codon usage pattern similar to that of highly expressed host genes (Sharp, Rogers, McConnell 1984). This suggests that concordance in codon usage between the host and the phage may be used as a proxy of phage codon adaptation. A simple measure of such concordance could be the correlation between host RSCU and phage RSCU, referred to hereafter as r_{RSCU}.

r_{RSCU} as a measure of phage codon adaptation has two problems. First, it can be increased not only by selection for codon adaptation, but also by biased mutation. For example, strongly AT-biased mutations shared by both the host and the phage will lead to a high r_{RSCU}. Such a high r_{RSCU} cannot be equated to a high degree of codon adaptation because adaptation, by definition, arises in response to selection. There is, however, one special case where r_{RSCU} can be reasonably used as a proxy of phage codon adaptation, and that is when we study phages parasitizing the same host and when the host has roughly equal nucleotide frequencies indicating unbiased mutations.

E. coli is approximately such a host species. Its genomic nucleotide frequencies are roughly equal, being 0.2462, 0.2541, 0.2537 and 0.2460 for nucleotides A, C, G and T, respectively. This indicates that mutations in *E. coli* do not lead to strong codon usage bias, in contrast to AT-biased or GC-biased mutations in many other bacterial species that can cause strong codon usage bias without any selection (Muto, Osawa 1987). Increasing the rate of unbiased mutations will lead to more randomized RSCU values and smaller r_{RSCU} values.

The benefit of using a host with equal genomic nucleotide frequencies (presumably resulting from unbiased mutation) is that the effect of tRNA-mediated selection is often unequivocally detectable. Table 2.1 illustrates *E. coli* codon usage of four codon families in

which tRNA-mediated selection favors A-, G-, C-, and U-ending codons, respectively. The most frequently used codon in each codon family matches the tRNA species with the highest gene copy numbers (Table 2.1). For example, there are four $tRNA^{Glu/UUC}$ genes forming Watson-Crick base pair with Glu codon GAA, but no $tRNA^{Glu/CUC}$. As tRNA gene copy number is well correlated with experimentally measured tRNA abundance (Percudani, Pavesi, Ottonello 1997), tRNA-mediated selection therefore should favor GAA, which is true (Table 2.1). What is remarkable is that this association between major codon and tRNA abundance is visible when tRNA-mediated selection favors A-, G-, C-, and U-ending codons, respectively (Table 2.1). If the *E. coli* genome had experienced strong AT-biased mutation, then tRNA-mediated selection for C-ending or G-ending codons may be invisible (i.e., A-ending and T-ending codons may still be the most frequently observed in spite of tRNA-mediated selection favoring C-ending and G-ending codons when AT-biased mutation dominates over the tRNA-mediated selection). For this reason, phages studied here are all *E. coli* phages.

The second problem with r_{RSCU} is that it does not capture all aspects of codon adaptation. This is illustrated in Table 2.2 which shows fictitious codon count and RSCU of highly expressed host genes and two phage genes (PG1 and PG2). RSCU values for codons in PG1 and PG2 are exactly the same, so r_{RSCU} for PG1 and PG2 will also be the same. However, PG2 is expected to be translated more efficiently than PG1 for the following reason. We notice that highly expressed host genes strongly avoid UUU in the Phe codon family (Table 2.2), suggesting that UUU cannot be translated efficiently by the host translation machinery. Given this, PG2 as a whole should be translated faster than PG1 because PG2 has only 90 "bad" UUU codons whereas PG1 has 180 "bad" UUU codons. In this case, the Gly codon family is "U-friendly" because an increased number of U-ending

codons will in fact improve translation. In contrast, the Phe codon family is "U-hostile" because increasing the number of U-ending codons will reduce translation efficiency. An ssDNA phage that cannot avoid high C→T mutations can nonetheless evolve codon adaptation by reducing the usage of codons in U-hostile codon families and increasing the usage of codons in U-friendly codon families as PG2 does (Table 2.2). This kind of adaptation is invisible to r_{RSCU}, but can be detected by CAI. We use the mean CAI value, computed from all genes in a phage genome with highly expressed host genes as a reference set, as an alternative measure of phage codon adaptation.

Phages are essentially a mosaic of genes sampled from a pool of frolicking phage genomes. For example, while many related tailed phages have nearly identical genome organization such as "DNA packaging-head-tail-tail fiber-lysis-lysogeny-DNA replication-transcription regulation" (Desiere et al. 2001), essentially any function in a phage can be fulfilled by one of many distinct genes with homologous function but little sequence similarity (Brussow, Kutter 2005). In other words, horizontal gene transfer is rampant in phage so that individual genes in each phage could differ dramatically in evolutionary history and codon usage. Consequently, a mean/median CAI may not be representative of all genes in a phage genome. For this reason we have added standard deviation of CAI values in the supplemental table (Table A.1) to show that the among-gene difference in CAI is actually quite small.

Table 2.1 - The effect of tRNA-mediated selection in *E. coli*, whose genomic sequence has equal nucleotide frequencies, presumably resulting from little mutation bias.

AA	Codon	N[(1)]	tRNA[(2)]	CP[(3)]
Glu	GAA	4683	4	A-ending
	GAG	1459	0	
Phe	UUC	2229	2	C-ending
	UUU	872	0	
Leu_4[(4)]	CUA	54	1	
	CUG	5698	4	G-ending
	CUC	541	1	
	CUU	357	0	
Arg_4[(4)]	CGA	34	0	
	CGG	33	1	
	CGC	1530	0	
	CGU	2995	3	U-ending

(1) number of codons in highly expressed *E. coli* genes compiled in the EMBOSS package (Rice, Longden, Bleasby 2000).
(2) number of *E. coli* tRNA genes with anticodon forming Watson-Crick pairing with the associated codon. Nucleotide A at the 1st anticodon position is mostly modified to inosine.
(3) CP: Codon preferred by tRNA
(4) Leu and Arg are coded by a four-codon subfamily and a two-codon subfamily. Leu_4 and Arg_4 refer to their respective four-codon subfamily.

Table 2.2 - Fictitious codon usage for highly expressed host genes (HOST) and and two phage genes (PG1 and PG2). r_{RSCU} between HOST and PG1 is identical to that between HOST and PG2, but PG2 will have higher CAI than PG1 when CAI is computed with HOST as the reference set of genes.

		Count			RSCU		
AA	Codon	HOST	PG1	PG2	HOST	PG1	PG2
Gly	GGA	400	50	75	0.8889	1	1
	GGG	300	30	45	0.6667	0.6	0.6
	GGC	100	20	30	0.2222	0.4	0.4
	GGU	1000	100	150	2.2222	2	2
Phe	UUC	2000	20	10	1.8182	0.2	0.2
	UUU	200	180	90	0.1818	1.8	1.8

2.3.2 Effect of phage-encoded tRNA genes on phage codon usage

Some phage genomes are long known to encode tRNA genes (Chattopadhyay, Ghosh 1988; Mandal, Ghosh 1988), e.g., Enterobacteria phage WV8 carries 20 tRNA genes on its genome. Phage-encoded tRNAs tend to have anticodons decoding codons overused in the phage genes but rarely used in host genes (Kunisawa 1992; Kunisawa 2000; Bailly-Bechet, Vergassola, Rocha 2007; Enav, Beja, Mandel-Gutfreund 2012). Such phage-encoded tRNAs would alter host tRNA pool, render the phage less dependent on the host tRNAs, and reduce the need (selection pressure) for the phage genes to evolve towards a codon usage pattern similar to that of the host genes. In other words, such tRNA genes would tend to reduce r_{RSCU} and CAI, and need to be taken into consideration in studying phage codon adaptation, especially in characterizing the difference between dsDNA and ssDNA phages because the latter do not encode tRNA genes in their genomes.

2.3.3 Effect of C→T mutation bias on codon usage of ssDNA phages

Mutation rate differs much between ssDNA and dsDNA phages. While dsDNA is well protected against mutation agents, ssDNA is subject to a high rate of DNA decay, especially spontaneous deamination leading to C→T mutations, the rate of which is about 100 times higher in ssDNA than in dsDNA (Frederico, Kunkel, Shaw 1990). The high mutation rate of ssDNA phages relative to dsDNA phages impact strongly on genomic GC content (Xia, Yuen 2005) and codon usage bias (Cardinale, Duffy 2011). For this reason, one would predict that, given the same tRNA-mediated selection for codon usage bias, dsDNA phages would achieve better codon adaptation than ssDNA phages.

2.3.4 Coevolution time and maximum r_{RSCU}

We have predicted that tRNA-mediated selection will increase r_{RSCU} and that increased mutation rate will decrease r_{RSCU} in *E. coli* phage. However, testing these predictions is confounded by coevolution time between phages and their host. Suppose a group of phages, given sufficient coevolution time with *E. coli,* would reach a maximum r_{RSCU}. When we sample these phage lineages, some may have coevolved sufficiently long to have reached the maximum r_{RSCU} while others may be far from reaching the maximum because they may have invaded *E. coli* only recently. Thus, both dsDNA and ssDNA phages may have some of their members with low r_{RSCU} values, but we predict that the maximum r_{RSCU} value should be much greater for dsDNA phages than for ssDNA phages.

In short, we predict that (1) for dsDNA phages, r_{RSCU} should decrease with the number of tRNA genes encoded by the phage genome, with phage-encoded tRNAs likely decoding codons over-used by phage mRNAs but rarely used by host mRNAs, (2) r_{RSCU} should be greater for dsDNA phages than ssDNA phages when the effect of phage-encoded tRNA

genes has been taken into consideration, and maximum r_{RSCU} should in particular be much greater for dsDNA phages than for ssDNA phages, and (3) ssDNA phages with a strong C→T mutation bias may evolve to increase the usage of codons in U-friendly codon families and reduce the usage of codons in U-hostile codon families. We report results confirming these predictions.

2.4 Materials and Methods

2.4.1 Genomic Data and Processing

The genome sequences of 469 dsDNA phages, 41 ssDNA phages and their corresponding bacterial hosts were downloaded from GenBank, of which 71 have *E. coli* specified as their host in the "/HOST" tag in "FEATURES" table, including 60 dsDNA phages and 11 ssDNA phages. All phage genomes were searched for encoded tRNAs by using tRNAscan-SE Search Server (Schattner, Brooks, Lowe 2005). The complete compilation with phage name, phage family, phage accession, strand, phage genome length, genomic GC%, number of coding sequences (CDSs) in each phage genome, genomic TC skew defined as $(N_T\text{-}N_C)/(N_T\text{+}N_C)$ where N_C and N_T are the genomic counts of nucleotides C and T, number of tRNA genes encoded in each phage genome, mean, median, standard deviation CAI and r_{RSCU} were included in a supplemental table (Table A.1).

E. coli has many strains sequenced, but the "/HOST" tag in most annotated viral genomes gives only species name (i.e., *E. coli*), with no strain-specific information. For this reason, the host GC% and RSCU are computed from the average of all *E. coli* genomes (The difference among *E. coli* strains is minimal). The mean *E. coli* genome length is 5024514 nt, mean number of CDSs is 4692.2 and mean genomic GC% is 50.68. The genomic accession

numbers of all *E. coli* strains used to compute the average statistics are also included in the supplemental table (Table A.1). The classification of phages into temperate and virulent categories is based on three publications (Lima-Mendez, Toussaint, Leplae 2007; Deschavanne, DuBow, Regeard 2010; McNair, Bailey, Edwards 2012).

2.4.2 Indices of codon adaptation

Coding sequences (CDSs) and tRNA genes in each phage and host genomes were extracted and RSCU computed by using DAMBE (Xia 2013b). r_{RSCU} (correlation between host and phage RSCU values) is taken as a measure of phage codon adaptation to the host translation machinery, with justifications outlined in the introduction. Single-codon families such as the Met (coded by AUG) and Trp (coded by UGG) were excluded from computing r_{RSCU} because the RSCU value is 1 for the two codons regardless of codon usage. CAI was computed with the improved implementation (Xia 2007) and highly expressed *E. coli* genes as the reference gene set. Throughout the text, the codon usage of highly expressed *E. coli* genes refers to the codon usage table compiled and distributed with the EMBOSS package (Rice, Longden, Bleasby 2000). The median CAI for protein-coding genes for each phage is used as an alternative measure of phage codon adaptation.

We did not use N_c (Wright 1990; Sun, Yang, X. 2013) as a measure of codon adaptation for the following reason. For an *E. coli* phage, selection by the host tRNA pool is expected to increase r_{RSCU} and CAI. In contrast, mutation, biased or not, will decrease r_{RSCU} and CAI. The effect of mutation and tRNA-mediated selection on N_c is more difficult to distinguish. In general, tRNA-mediated selection will decrease N_c, but biased mutation will also decrease N_c. For this reason, N_c is not good for measuring codon adaptation in *E. coli* phages.

2.5 Results

Twenty-two dsDNA phage species encode tRNA genes in their genomes (13 from Myoviridae, four from Podoviridae, and five from Siphoviridae; Supplemental table (Table A.1)), whereas none of the ssDNA phage genomes carry tRNA genes. Before making comparisons in codon usage between dsDNA and ssDNA phages, it is important to test if phage-encoded tRNA genes can affect codon usage. The presence of an effect implies that the fair comparison should only be carried out between ssDNA phages and those dsDNA phages that do not carry tRNA genes.

2.5.1 Effect of phage-encoded tRNA on codon adaptation in dsDNA phage

We have reasoned before that phage-encoded tRNA genes may reduce r_{RSCU}, especially if these tRNAs tend to decode codons overused in the phage genes but underused in host genes. There is indeed a highly significant ($p < 0.0001$) negative relationship between r_{RSCU} and the number of tRNA genes encoded in the phage genome (Figure 2.1). The use of an exponential decay to fit the negative relationship is based on the rationale that, if the number of tRNA genes in the phage approaches infinity, then the codon usage of the phage would approach complete independence of the host tRNA pool, with r_{RSCU} approaching zero. A significant ($p = 0.0260$) negative relationship is also observed between CAI and the number of tRNA genes encoded in the phage genome.

What tRNA genes would benefit dsDNA phages that carry them? Translation of codons that are overused in phage genes but decoded by few host tRNAs would benefit from having extra cognate tRNAs from the phage genomes. Take R-ending codon for example (where R stands for purine). If the host tRNA pool favors G-ending codon, but A-ending codon is overused by phage genes, then it is beneficial for the phage to carry tRNA genes with a

wobble U to decode the overused A-ending codons. Similarly, if the host has few tRNAs decoding G-ending codons and uses few G-ending codons, but the phage uses many more G-ending codons, then it would be beneficial for phage tRNAs to have a wobble C to decode its relatively more frequently used G-ending codons.

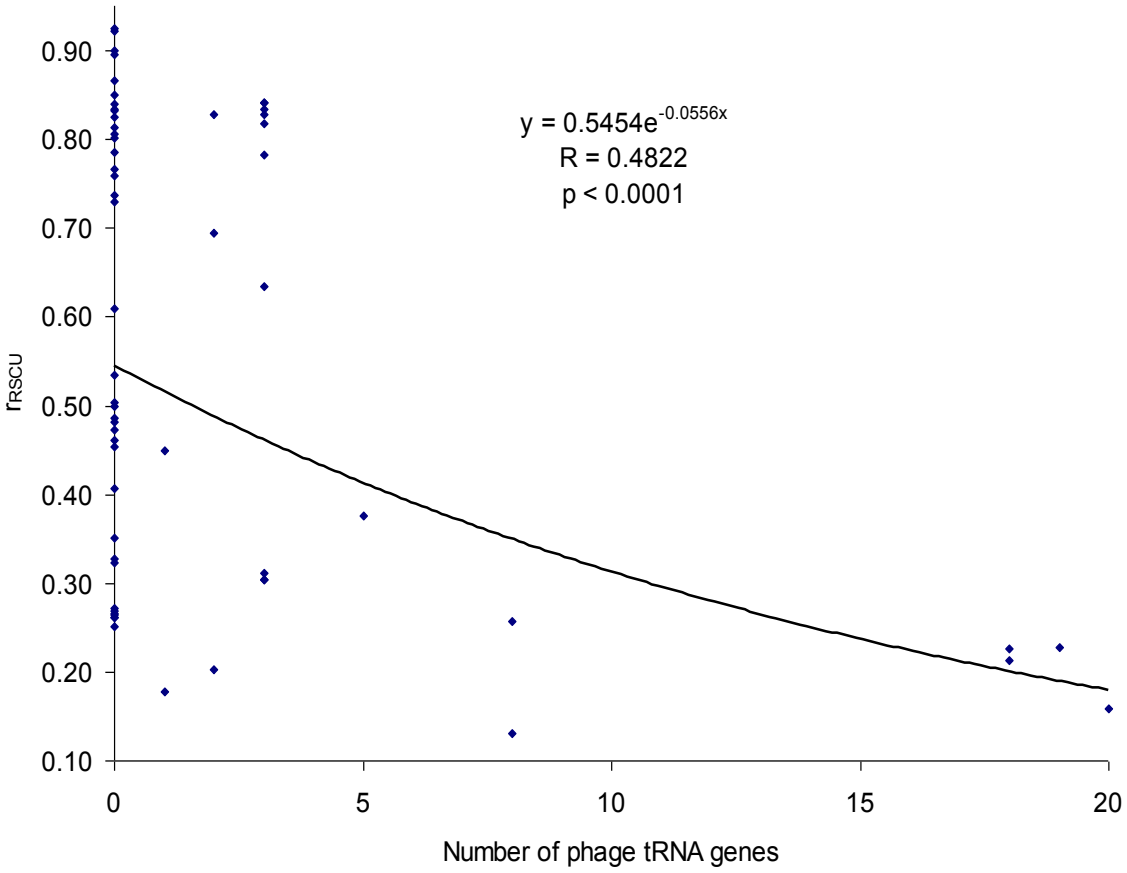

Figure 2.1 - Codon adaptation of the phage genes, measured by r_{RSCU}, decreases with increasing number of tRNA genes encoded in phage genomes.

Three general rules can be derived from the results in Table 2.3 which shows the R-ending codon usage of highly expressed *E. coli* genes and two dsDNA phages each carrying a set of tRNA genes. First, if phage codon usage bias is the same as that of *E. coli* (e.g., GAR, AAR and AGR codons for amino acids E, K, and R, respectively), then the phage-

encoded tRNAs will decode the most frequently used codon. Second, if phage codon usage bias is opposite to that of the host (e.g., GGR, UUR, CCR, and UCR codons for amino acids G, L, P and S, respectively), then the phage-encoded tRNAs will decode the codon overused in the phage but underused in the host. Third, if phage genes use the two R-ending codons roughly equally (e.g., CAR codons for amino acid Q), then the phage may carry tRNAs for both codons. While only two phage species are included in Table 2.3, the three rules are shared among other phage species with phage-encoded tRNAs.

The three rules are generally consistent with the interpretation that phage-encoded tRNAs facilitate translation of phage mRNAs. Similar findings, but less complete, have also been reported in previous studies on T4-like phages (Kunisawa 1992; Bailly-Bechet, Vergassola, Rocha 2007; Enav, Beja, Mandel-Gutfreund 2012). They are also consistent with previous experiments in which alteration of *E. coli* tRNA pool is associated with changed translation efficiency of transgenes (Kleber-Janke, Becker 2000).

One may note that Table 2.3 includes only R-ending codons. Can we extend the pattern to Y-ending codons (where Y stands for pyrimidine)? Suppose that the host overuses C-ending codons, with many tRNAs with a wobble G, but the phage overuses U-ending codons. Should we not predict that phage genomes should encode tRNAs with a wobble A to decode its overused U-ending codons? However, this prediction cannot be tested because a tRNA with wobble A would interfere with translation. That is, once such a tRNA is in the P-site, it interferes with the tRNA at the A-site (Lim 1994). Thus, Y-ending codons are decoded by either tRNAs with a wobble G or tRNA with a wobble A-derived inosine. This was overlooked in a previous study on tRNAs encoded in phage T4 (Kunisawa 1992).

Table 2.3 - Number of A- or G-ending codons (N_{cod}), relative synonymous codon usage (RSCU) and number of tRNA genes (N_{tRNA}) for *E. coli* and two phage species (WV8 and bV_EcoS_AKFV33). See text for reasons of including only R-ending codons.

		E. coli(1)			WV8			bV_EcoS_AKFV33		
AA	Codon	N_{cod}	RSCU	N_{tRNA}	N_{cod}	RSCU	N_{tRNA}	N_{cod}	RSCU	N_{tRNA}
E	GAA	4683	1.5249	4	1125	1.259	1	1489	1.365	1
E	GAG	1459	0.4751		662	0.741		692	0.635	
G	GGA	118	0.0679	1	245	0.584	1			
G	GGG	267	0.1536	1	150	0.357				
K	AAA	4129	1.5945	5	1262	1.195	1	1551	1.364	1
K	AAG	1050	0.4055		851	0.805	1	723	0.636	1
L	CUA	54	0.0325	1	233	0.745	1	544	1.335	1
L	CUG	5698	3.4274	3	318	1.017		433	1.063	
L	UUA	210	0.7735	1				718	1.453	1
L	UUG	333	1.2265	1				270	0.547	
P	CCA	474	0.5636	1	408	2.032	1	428	1.558	1
P	CCG	2509	2.9834	1	62	0.309		154	0.561	
Q	CAA	550	0.3551	2	481	1.058	1	593	1.06	1
Q	CAG	2548	1.6449	2	428	0.942	1	526	0.94	1
R	AGA	21	1.2353	8	438	1.581	1	317	1.461	1
R	AGG	13	0.7647	1	116	0.419		117	0.539	
S	UCA	189	0.2612	1	498	1.64	1			
S	UCG	275	0.3801	1	38	0.125				
T	ACA	181	0.16	1				447	1.002	1
T	ACG	526	0.4649	1				164	0.368	
V	GUA	1329	0.8047	5				765	1.508	1
V	GUG	1784	1.0802					231	0.455	

(1) From highly expressed *E. coli* genes, as compiled in the EMBOSS distribution (Rice, Longden and Bleasby 2000).

2.5.2 Difference in r_{RSCU} between dsDNA and ssDNA phages

Given the significant effect of phage-encoded tRNA on r_{RSCU} (Figure 2.1 and Table 2.3), all phage genomes with encoded tRNA genes were excluded in all comparisons between dsDNA phages and ssDNA phages because none of the ssDNA phage genomes encode tRNA genes. This leaves 38 dsDNA phages and 11 ssDNA phages for further comparisons in r_{RSCU}.

r_{RSCU} is significantly greater for dsDNA phages than for ssDNA phages (0.5917 for the former and 0.3273 for the latter, t= 3.6533, DF = 47, p = 0.0008, Table 2.4). To test if it is the C→T biased mutation that is chiefly responsible for the reduced r_{RSCU} values for the ssDNA phages, we computed the r_{RSCU} values separately for the R-ending codons and Y-ending codons (Table 2.5). The r_{RSCU} values for the R-ending codons ($r_{RSCU.R}$) are significantly greater than those for the Y-ending codons ($r_{RSCU.Y}$), with the mean being 0.5217 for $r_{RSCU.R}$ and 0.1074 for $r_{RSCU.Y}$ (Table 2.5). The difference is highly significant (paired-sample t-test: t = 17.2872, DF = 10, p < 0.0001), assuming data independence.

Table 2.4 - Mean and distribution of r_{RSCU} values for various dsDNA and ssDNA phage families.

Type	Phage family	n	Minimum	Maximum	Average	SD
dsDNA	Myoviridae	9	0.3437	0.9207	0.6953	0.2359
	Podoviridae	12	0.2553	0.8034	0.4216	0.1859
	Siphoviridae	16	0.2412	0.8955	0.66	0.2355
	Tectiviridae	1	0.6084	0.6084	0.6084	N/A
ssDNA	Inoviridae	4	0.27	0.3922	0.3449	0.0524
	Microviridae	7	0.2757	0.3709	0.3173	0.0409

Table 2.5 - Contrasting r_{RSCU} values for R-ending codons and for Y-ending codons (designated by $r_{RSCU.R}$ and $r_{RSCU.Y}$, respectively).

Family	ACCN	$r_{RSCU.R}$	$r_{RSCU.Y}$
Microviridae	NC_001330	0.6504	0.0854
Microviridae	NC_001420	0.453	0.0332
Microviridae	NC_007856	0.4652	0.0447
Microviridae	NC_007817	0.4168	0.02
Microviridae	NC_001422	0.4497	0.0843
Microviridae	NC_012868	0.6009	0.1118
Microviridae	NC_007821	0.603	0.1158
Inoviridae	NC_001332	0.5475	0.1709
Inoviridae	NC_001954	0.4753	0.2154
Inoviridae	NC_002014	0.5892	0.2105
Inoviridae	NC_003287	0.4876	0.0894
Mean		0.5217	0.1074

Because some phages may not have enough time coevolving with their host, their r_{RSCU} may not have reached the maximum possible. For example, if a dsDNA phage has recently switched to a host with a different codon usage pattern, then we would not expect it to have a high r_{RSCU} value because codon adaptation takes time to evolve. However, given enough time, we expect dsDNA phages to reach a higher r_{RSCU} than ssDNA phages whose mutation rate is higher than that of dsDNA phages. The mean and distribution of r_{RSCU} values for the dsDNA and ssDNA phage (Table 2.4) is consistent with this interpretation. The maximum r_{RSCU} observed is only 0.3922 for ssDNA phages, but 0.9207 for dsDNA phages

(Enterobacteria phage Mu in Myoviridae). The mean and standard variation of r_{RSCU} values for ssDNA phage is 0.3273 and 0.0450, respectively, so that the probability of having an r_{RSCU} value as large as 0.5 is less than 0.0001 for ssDNA phages.

When a phage species has a small r_{RSCU} value, it could either be due to weakened selection (e.g., the phage carries a large number of its own tRNA genes), strong mutation pressure disrupting codon adaptation, or insufficient coevolution time. Given that the three dsDNA phage families and the two ssDNA phage families all have multiple phage lineages parasitizing *E. coli,* we may assume that the phages should have coevolved with *E. coli* for sufficiently long time for codon adaptation to reach a mutation-selection equilibrium. Also, the comparison above between the dsDNA and ssDNA phages excluded phages with phage-encoded tRNA genes, so all these phages should have experienced roughly the same host tRNA-mediated selection. The most plausible explanation for the difference in r_{RSCU} between the dsDNA and ssDNA phages is the higher mutation pressure in ssDNA phages that disrupt codon adaptation.

2.5.3 Effect of life cycle (temperate vs. virulent) on r_{RSCU} in dsDNA phages

dsDNA phages differ in their life cycles, some being temperate with a lysogenic phage and some are virulent with only lytic phase, although lysogenic phages can become lytic through mutations at lysogenic conversion genes (van Vliet et al. 1978; Brussow, Kutter 2005). Temperate phages are expected to have better concordance in codon usage with the host (i.e., higher r_{RSCU} values) than lytic phages for two reasons. First, a prophage and its lysogen share the same mutation spectrum as the host DNA. Second, they have increased chance of recombining with or acquiring host genes or gene segments. For example, phage λ

and phage μ carry a piece of host genome when they switch from the lysogenic phase to the lytic phase.

The expectation is borne out by empirical data (Table 2.6), with r_{RSCU} significantly greater in temperate phages than in virulent phages with two-sample t-tests ($DF = 7$, $t = 11.5914$, $p < 0.0001$ for Myoviridae; $DF = 9$, $t = 5.7328$, $p = 0.0003$ for Podoviridae, $DF = 12$, $t = 10.4545$, $p < 0.0001$ for Siphoviridae). A two-way ANOVA accounts for 91.24% of total variance in r_{RSCU}, with r_{RSCU} differing highly significantly between temperate and virulent phages ($F = 280.9918$, $DF_{model} = 1$, $DF_{error} = 28$, $p < 0.0001$), significantly among the three dsDNA phage families ($F = 5.095$, $DF = 2$, $p = 0.0130$), but with no significant interaction ($F = 0.2101$, $DF = 2$, $p = 0.81175$).

Table 2.6 - Effect of life cycle of dsDNA phages on codon usage concordance between phage and host, measured by rRSCU. The phages are organized by phage families (PhageFam) and then by life cycle (LifeCycle: temperate or virulent) within each family.

PhageFam	PhageName	Accession	LifeCycle	r_{RSCU}
Myoviridae	Enterobacteria phage Mu	NC_000929	Temperate	0.9207
Myoviridae	Enterobacteria phage P2	NC_001895	Temperate	0.9011
Myoviridae	Enterobacteria phage P4	NC_001609	Temperate	0.8287
Myoviridae	Enterobacteria phage SfV	NC_003444	Temperate	0.875
Myoviridae	Escherichia phage D108	NC_013594	Temperate	0.9207
Myoviridae	Enterobacteria phage JSE	NC_012740	Virulent	0.4789
Myoviridae	Enterobacteria phage Phi1	NC_009821	Virulent	0.4971
Myoviridae	Enterobacteria phage phiEcoM-GJ1	NC_010106	Virulent	0.3437
Myoviridae	Enterobacteria phage RB49	NC_005066	Virulent	0.4917
Podoviridae	Escherichia phage phiV10	NC_007804	Temperate	0.7308
Podoviridae	Stx2 converting phage I	NC_003525	Temperate	0.8034
Podoviridae	Enterobacteria phage 13a	NC_011045	Virulent	0.3181
Podoviridae	Enterobacteria phage EcoDS1	NC_011042	Virulent	0.4021
Podoviridae	Enterobacteria phage K1-5	NC_008152	Virulent	0.2629
Podoviridae	Enterobacteria phage K1E	NC_007637	Virulent	0.2553
Podoviridae	Enterobacteria phage K1F	NC_007456	Virulent	0.2553

Continued on next page

Continued from previous page

PhageFam	PhageName	Accession	LifeCycle	r_{RSCU}
Podoviridae	Enterobacteria phage N4	NC_008720	Virulent	0.2661
Podoviridae	Enterobacteria phage T3	NC_003298	Virulent	0.5306
Podoviridae	Enterobacteria phage T7	NC_001604	Virulent	0.3274
Podoviridae	Enterobacteria phage BA14	NC_011040	Virulent	0.4504
Siphoviridae	Enterobacteria phage BP-4795	NC_004813	Temperate	0.8049
Siphoviridae	Enterobacteria phage cdtI	NC_009514	Temperate	0.8307
Siphoviridae	Enterobacteria phage HK022	NC_002166	Temperate	0.7416
Siphoviridae	Enterobacteria phage HK97	NC_002167	Temperate	0.7303
Siphoviridae	Enterobacteria phage lambda	NC_001416	Temperate	0.852
Siphoviridae	Enterobacteria phage N15	NC_001901	Temperate	0.8955
Siphoviridae	Escherichia Stx1 converting bacteriophage	NC_004913	Temperate	0.8108
Siphoviridae	Stx2-converting phage 1717	NC_011357	Temperate	0.8335
Siphoviridae	Enterobacteria phage SSL-2009a	NC_012223	Temperate	0.7853
Siphoviridae	Enterobacteria phage EPS7	NC_010583	Virulent	0.2583
Siphoviridae	Enterobacteria phage JK06	NC_007291	Virulent	0.2565
Siphoviridae	Enterobacteria phage RTP	NC_007603	Virulent	0.2412
Siphoviridae	Enterobacteria phage T1	NC_005833	Virulent	0.4637
Siphoviridae	Enterobacteria phage TLS	NC_009540	Virulent	0.4734

2.5.4 A new type of codon adaptation mediated by C→T biased mutation

Some ssDNA phages have strong C→T mutations as measured by $SKEW_{TC}$ defined as

$$SKEW_{TC} = \frac{N_T - N_C}{N_T + N_C} \quad (1)$$

where N_T and N_C are the count of nucleotides T and C, respectively. $SKEW_{TC}$ is expected to increase with increased C→T mutation rate and result in overuse of U-ending codons. For example, Enterobacteria phage Ike (NC_002014, Inoviridae) has a $SKEW_{TC}$ value of 0.2893, with U-ending codons being the most frequent in all Y-ending or N-ending codon families. The effect of biased mutation on codon usage has also been shown for several other ssDNA phages (Cardinale, Duffy 2011). This bias in favor of U-ending codons interferes with codon

adaptation because *E. coli* translation machinery does not favor U-ending codons in most codon families. Highly expressed *E. coli* genes, as compiled in the EMBOSS distribution (Rice, Longden, Bleasby 2000) or in Ran and Higgs (2012), have U-ending codons being the most frequent in four codon families, i.e., Gly, Arg_4 (the CGN codon subfamily for Arg), Ser_4 (the UCN codon subfamily for Ser) and Val. Take the Val (GUN) codon family for example. The RSCU values for GUA, GUC, GUG, GUU are 0.8047, 0.4989, 1.0802, and 1.6161, respectively, based on the EMBOSS distribution (Rice, Longden, Bleasby 2000). Such a codon family is "U-friendly" because U-ending codons are preferred and C→T biased mutation will consequently improve translation elongation. In contrast, the other codon families containing U-ending codons have C-ending codons more frequent than U-ending codons based on the highly expressed *E. coli* protein-coding genes. These codon families will be designated as U-hostile. T-biased mutation in ssDNA phages would enhance codon adaptation in the four U-friendly codon families, but would go against codon adaptation in the U-hostile codon families.

What can ssDNA phages do to increase their translation elongation efficiency in face of the C→T mutation? One obvious solution to the problem is illustrated in Table 2.2 with codon frequencies of two codon families (Gly and Phe) from two fictitious phage genes (designated as PG1 and PG2, respectively) and from the host. We can infer U-friendliness of the host translation machinery based on codon usage of host genes. The Gly codon family is U-friendly, with the host machinery strongly preferring U-ending codons. The Phe codon family is U-hostile with host translation machinery strongly favoring C-ending codons (Table 2.2). The total number of codons for the two genes is the same and equal to 400, and the RSCU for each codon is also identical for two genes (Table 2.2). Thus, r_{RSCU} between

PG1 and host would be exactly the same as that between PG2 and host. However, we note that the PG2 could be translated more efficiently than PG1 because the former has only 90 "bad" UUU codons whereas the latter has 180. This differential translation elongation efficiency is not reflected by RSCU but is by CAI. For example, with the data in Table 2.2 and assuming no other codons except for those listed in Table 2.2, we have CAI being 0.2577 for PG1 but 0.3686 for PG2 when host codon frequencies are used as the reference set.

The example illustrated above suggested that *E. coli* ssDNA phages with strong C→T mutation bias can improve their translation elongation efficiency by overusing the codons in the four U-friendly codon families and decreasing the codons in the U-hostile codon families. This leads to the prediction that the summed frequencies of codons in the four U-friendly codon families, designated as F_4, should increase with $SKEW_{TC}$. That is, when U-ending codons are increased by U-biased mutations, these U-ending codons should be more concentrated in the four U-friendly codon families. This prediction is strongly supported by data from the 11 ssDNA *E. coli* phages (Figure 2.2), with the correlation between F_4 and $SKEW_{TC3} = 0.707$ ($p = 0.0151$). Furthermore, F_4 is significantly and positively correlated with mean CAI from the 11 ssDNA phages ($r = 0.6595$, $p = 0.0273$). The result in Figure 2.2 is consistent with the interpretation that increased C→T mutation drives the increased use of codons in the four U-friendly codon families. Thus, while the ssDNA phages cannot fight against the C→T mutation, they have evolved to minimize the disruptive effect of this biased mutation on codon adaptation by coding more amino acids in the four U-friendly codon families.

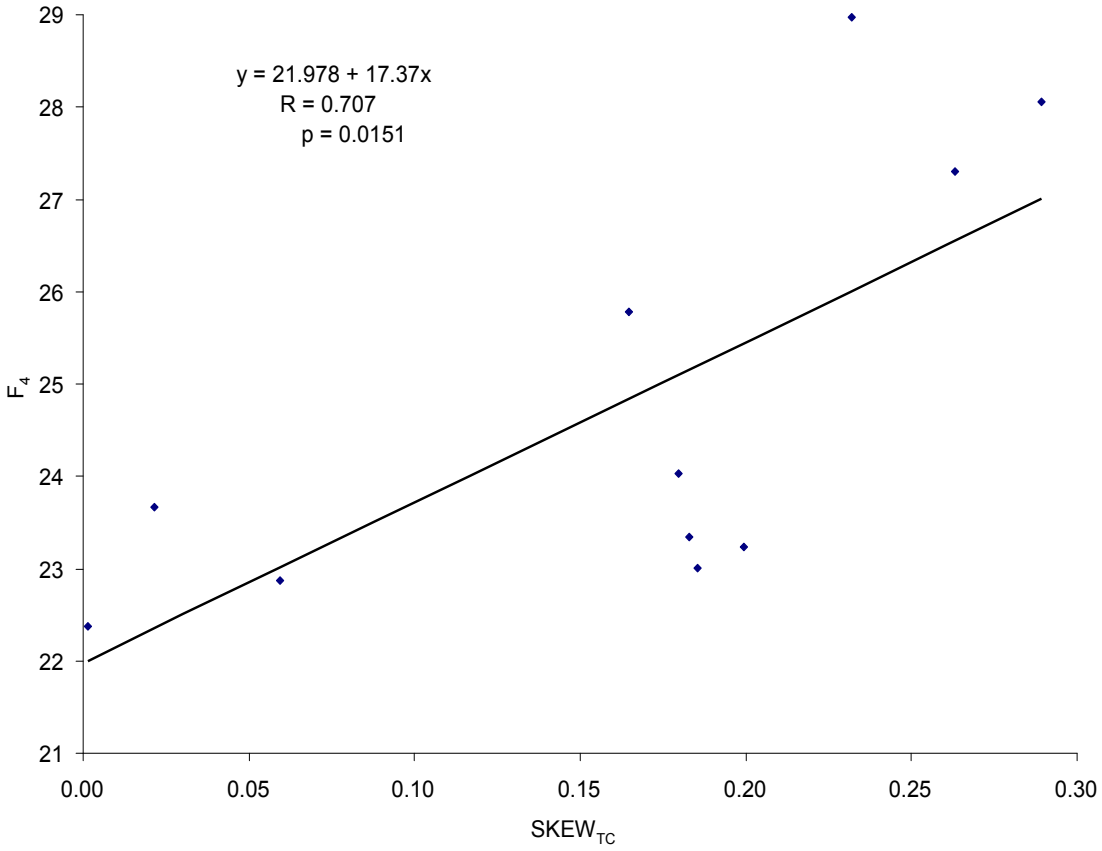

Figure 2.2 - Positive association between $SKEW_{TC}$, defined as $(N_T - N_C)/(N_T + N_C)$ where N_i is the number of nucleotide *i* in a phage genome, and F_4, total number of codons in four codon families (Gly, Arg_4, Ser_4 and Val) in which highly expressed *E. coli* genes prefer U-ending codons against C-ending codons. Results are from 11 ssDNA *E. coli* phages.

The usage of Ser codons for Enterobacteria phage Ike (NC_002014, Inoviridae) illustrates this special codon adaptation well. Ser is coded by the four-codon UCN and the two-codon AGY codon subfamilies. In the AGY codon subfamily, highly expressed *E. coli* genes prefer AGC against AGU, suggesting that AGU is a "bad" codon. C→T mutations will lead to many "bad" AGU codons if Ser is largely encoded by the AGY subfamily. In contrast, in the UCN subfamily, highly expressed *E. coli* genes strongly prefer UCU against other synonymous codons, suggesting that UCU is a "good" codon. C→T mutations will lead to many "good" UCU codons if Ser is largely encoded by the UCN subfamily. In this

conceptual framework, it is easy to understand that 88.4% of Ser codons in Enterobacteria phage Ike belong to the UCN subfamily. Because of this adaptive trick, the mean CAI value for ssDNA phages is almost as large as that for dsDNA phages (0.4768 for dsDNA phages and 0.4743 for ssDNA phages, excluding the 22 phages with phage-encoded tRNA genes), with no statistically significant difference.

The type of codon adaptation outlined above, i.e., by switching codon usage from U-hostile codon families to U-friendly codon families, implies increased nonsynonymous substitution with increased C→T mutation. A simple way to check this is to test the change of UUC and CCN frequencies with increased C→T mutation rate. We used TC Skew at the third codon position ($SKEW_{TC3}$) to measure C→T mutation and checked how the frequencies of UUN and CCN codons would change $SKEW_{TC3}$. The frequency of UUN codons increases ($p = 0.0008$, Figure 2.3), and that of CCN codons decreases ($p = 0.0320$, Figure 2.3), with increasing $SKEW_{TC3}$, consistent with the expectation. However, the sharp increase in UUN codons and the relatively slow decrease in CCN codons (Figure 2.3) suggest that the increase in UUN codon is not entirely due to the decrease of CCN codons. Similar response of nonsynonymous mutation rate to directional mutation pressure has also been documented in several other studies (Sueoka 1961; Lobry 2004; Urbina, Tang, Higgs 2006).

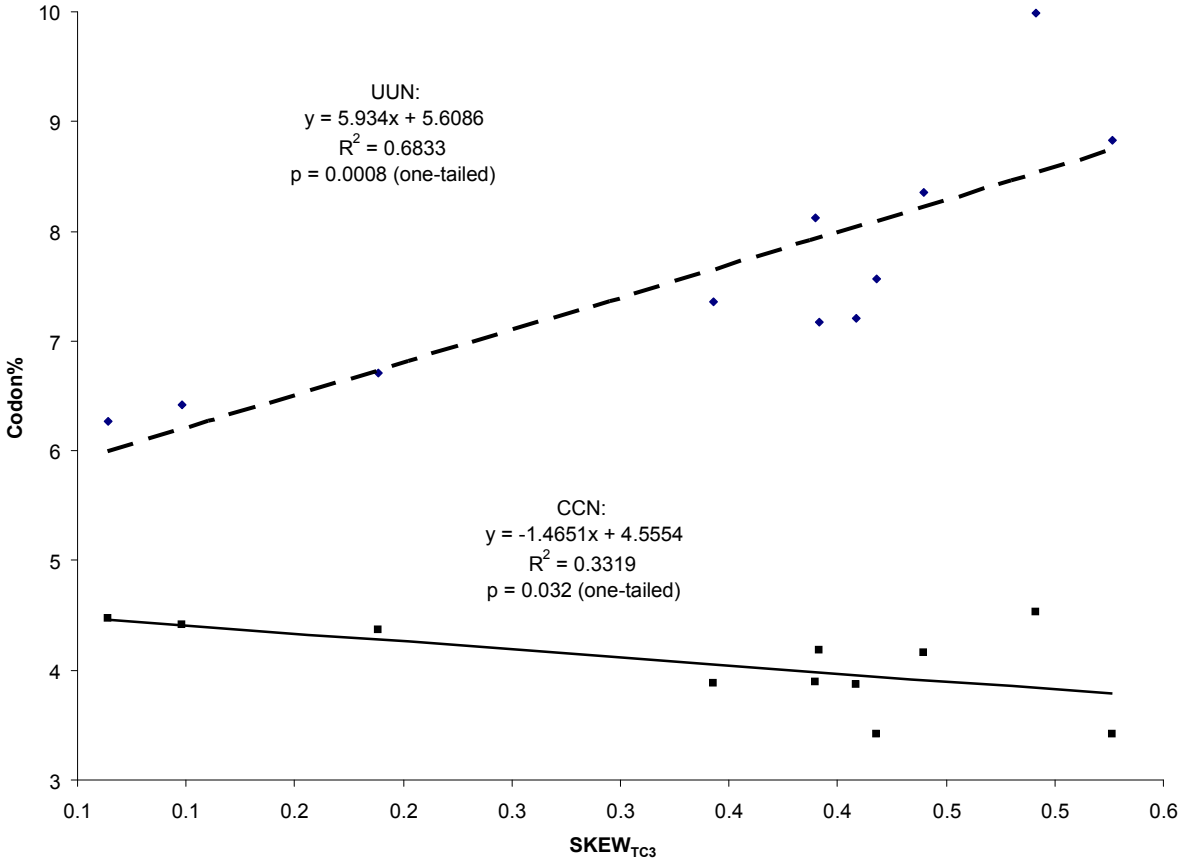

Figure 2.3- UUN codons increases, and CCN codons decreases, with C→T mutation measured by TC skew at the third codon position ($SKEW_{TC3}$), but at different extent.

The results above suggest to us that our empirical test of the new type of codon adaptation in Figure 2.2 is incorrect. For example, the Val codon family (coded by GUN) is U-friendly and its usage increases with C→T mutation bias, thus supporting the prediction from the hypothesized new type of codon adaptation. However, the increase may have nothing to do with codon adaptation, but may be simply due to the increase of all U-containing codons and the decrease of C-containing codons with increasing C→T mutation bias. Thus, only codon families that do not contain C or U at the first and second codon positions are relevant to test the prediction of a positive association between the usage of U-friendly codon families and $SKEW_{TC3}$. Among the U-friendly codon families, only the Gly

codon family (coded by GGN) fulfills these criteria. The hypothesis is still supported as the percentage of Gly codons increased with $SKEW_{TC3}$ (r = 0.6837, p = 0.0204).

2.6 Discussion

Studying codon adaptation in phage is important not only in understanding the biology of translation, but also in practical applications. Several phages have been used to remove infectious biofilms (Azeredo, Sutherland 2008; Gladstone, Molineux, Bull 2012), to deliver vaccines (Clark, March 2004), or to treat human infections (Sau et al. 2005; Ranjan, Vidyarthi, Poddar 2007; Sau 2007; Skurnik, Pajunen, Kiljunen 2007; Goodridge 2010; Timms et al. 2010; Abedon et al. 2011), especially those caused by bacterial pathogens that have developed resistance to antibiotics. However, many of these phages do not have optimal codon usage for efficient replication. Studying codon adaptation in phages contributes to the theoretical foundation for re-engineering more efficient phages for therapeutic or industrial purposes (Skiena 2001). A database has been created to facilitate the study of phage codon adaptation to their hosts (Hilterbrand, Saelens, Putonti 2012).

2.6.1 Phage-encoded tRNA affect phage codon usage

We found that the number of tRNA genes carried by dsDNA phage genomes reduced the need for the phages to evolve a codon usage pattern similar to that of their hosts and that these phage-encoded tRNA facilitate the translation of overused phage codons, especially when the host provides few tRNAs for these phage codons (Figure 2.1 and Table 2.3). Several viral species have been found to alter host tRNA pool to favor the translation of the viral genes. HIV-1 viruses selectively enrich rare host tRNAs to decode A-ending codons overused in HIV-1 genes but rarely used by host genes (van Weringh et al. 2011), and such

selective enrichment has also been found in vaccinia and influenza A viruses (Pavon-Eternod et al. 2013).

Translation efficiency is sensitive to the change of tRNA pool (Kleber-Janke, Becker 2000). A gain/loss of a $tRNA^{Met/UAU}$ gene has resulted significant change in AUA codon frequencies, in both bivalve mitochondria and tunicate mitochondria (Xia et al. 2007; Xia 2012b). All these findings on the association of tRNA pool and codon usage suggest that translation efficiency of a target gene can not only be improved by optimizing the codon usage of the target gene, but also by modifying the tRNA pool where the target gene is translated. This latter approach has the advantage over the former because the former sometimes will alter the structure of the mRNA leading to reduced translation initiation efficiency (Kudla et al. 2009).

Phage-encoded tRNA genes provide phages with the opportunity to parasitize hosts with different codon usage and may therefore increase their host diversity (Sau et al. 2007). However, existing data do not allow the characterization of phage-encoded tRNA and host diversity because few phage species have their host diversity characterized. One way to characterize host diversity is by subjecting phages to a diverse array of hosts and checking for lytic activities (Villegas et al. 2009). Unfortunately, few such studies have been carried out.

2.6.2 Mutation plays a significant role in phage codon adaptation

The rate of spontaneous deamination leading to C→T mutation is about 100 times higher in ssDNA than in dsDNA (Frederico, Kunkel, Shaw 1990). These high C→T mutations prevent ssDNA phages from evolving a codon usage pattern as close to that of the host as dsDNA phages. This is substantiated by the observation that r_{RSCU} for R-ending

codons are significantly greater than r_{RSCU} for Y-ending codons in ssDNA phages (Table 2.5).

While our result is consistent with the mutation hypothesis, the lack of selection for Y-ending codons may also play a role in the poor concordance in RSCU for Y-ending codons between ssDNA phages and *E. coli.* A previous study (Xia 2008) strongly suggests that tRNAs with a wobble G are equally efficient in decoding C-ending and U-ending codons. This implies that C→T mutations will not be counterchecked by selection, leaving the ratio of U-ending to C-ending codons entirely to the mercy of mutation bias.

2.6.3 A new type of codon adaptation in ssDNA phage in response to the C→T mutation pressure

The C→T mutation pressure has driven ssDNA phages to evolve a previously unknown type of codon adaptation by biased usage of codon families. That is, they overuse U-friendly codon families in which C→T biased mutations improve codon adaptation, and avoid U-hostile codon families in which the biased mutation hampers codon adaptation (Figure 2.2). We have illustrated this adaption strategy with the codon usage in the Ser codon family for Enterobacteria phage Ike (NC_002014, Inoviridae) with a strong $SKEW_{TC}$ indicating a strong C→T mutation bias. This simple strategy allows the protein-coding genes in ssDNA phages to have CAI values comparable to those of dsDNA phages.

We have noticed an analogous codon adaptation in the six-codon Leu, Arg and Ser compound codon families in the yeast, *Saccharomyces cerevisiae,* in which the number of tRNA genes differ much between the four-codon subfamily and the two-codon subfamily. The yeast genome has 17 $tRNA^{Leu}$ genes for the two-codon UUR subfamily but only four $tRNA^{Leu}$ genes for the four-codon CUN codon family. The UUR codons account for 84% of

Leu codons in highly expressed yeast genes compiled in the EMBOSS distribution (Rice, Longden, Bleasby 2000). A similar pattern is observed for the Arg codon family. There are 16 $tRNA^{Ser}$ genes for the four-codon UCN subfamily and only two for the two-codon AGY codon subfamily. As expected, the UCN codons account for 89% of all Ser codons in highly expressed yeast genes. In short, whenever possible, selection for increased translation efficiency would drive protein-coding genes to maximize the use of codons that have many tRNAs to decode them.

Our study can be advanced in two ways. First, it should take into consideration the role of translation initiation in addition to translation elongation. Genes with poor translation initiation are not expected to increase their protein production with optimized codon usage. It is only genes with efficient translation initiation that are expected to increase protein production with improved codon-anticodon adaptation (Tuller et al. 2010).

Second, the existing phage genomic sequences still do not allow the construction of a sufficiently large phylogeny for phylogeny-based comparisons (Felsenstein 1985; Xia 2013b), mainly due to (1) the rapid evolution of phage genomes, especially ssDNA phage genomes, and (2) few homologous genes identifiable among phage species parasitizing *E. coli*. However, one could argue that, given the rapid evolutionary erosion of coancestry among these phage lineages, the data from different phage lineages may indeed be considered nearly independent. Phages are essentially a mosaic of genes sampled from a pool of frolicking phage genomes. For example, while a number of "related" tailed phages have nearly identical genome organization at function level such as "DNA packaging-head-tail-tail fiber-lysis-lysogeny-DNA replication-transcription regulation" (Desiere et al. 2001), essentially any function in a phage can be fulfilled by one of many distinct genes with "homologous" function but little sequence homology (Brussow, Kutter 2005). In other

words, horizontal gene transfer is so rampant that, coupled with rapid evolution, phylogenetic reconstruction based on sequence homology is nearly impossible. For example, a large number of phages have DNA polymerase, but these DNA polymerases apparently belong to a number of nonhomologous classes.

The difficulty in building a reliable phage tree also prevents an interesting question to be addressed. The loss/gain of tRNA genes may be related to host tRNA pool. Take AAR (Lys) codon family for example. If a phage species overusing AAA codons originally parasitizes a host overusing AAG codons and having abundant $tRNA^{Lys/CUU}$ but rare $tRNA^{Lys/UUU}$, then the phage would benefit from retaining a $tRNA^{Lys/UUU}$ gene decoding its overused AAA codons. If the phage subsequently switched to a host overusing AAA codons and having abundant $tRNA^{Lys/UUU}$, then the phage-encoded $tRNA^{Lys/UUU}$ gene would be of little value and would be prone to gene loss. Addressing such a question would be straightforward if one can build a reliable phage tree so that the gain/loss of tRNA genes can be mapped onto the tree.

3 Chapter Three

Characterizing the effect of mutation and selection on codon adaptation in *Escherichia coli* phage

3.1 Abstract

Codon adaptation in phage directly impacts translation efficiency and the utility of phage in controlling bacterial population in health and industry. To evaluate the effect of mutation and selection on phage codon usage, we developed an index (φ) to measure selection imposed by host translation machinery, based on the difference in codon usage between all host genes and highly expressed host genes. We developed linear and nonlinear models to estimate the C→T mutation bias in different phage lineages and to evaluate the relative effect of mutation and host selection on phage codon usage. C→T biased mutations affect not only synonymous codon usage, but also nonsynonymous substitutions at second codon positions, especially in ssDNA phages. Selection from host translation machinery (φ) affects codon adaptation in both dsDNA and ssDNA phages. dsDNA phages without phage-encoded tRNA genes tend to have better codon adaptation than those with tRNAs. Strand asymmetry with the associated local variation in mutation bias can significantly interfere with codon adaptation in both dsDNA and ssDNA phages. If mutation is strongly U-biased in half of the phage genome and C-biased in the other half, then preference of C-ending codons by host translation machinery would be efficient in only half of the phage genome.

3.2 Contributions

The data, results and interpretations in this chapter were submitted to *Genetics* journal. Shivapriya Chithambaram (SC) is the first author, Ramanandan Prabhakaran (RP) is the co-author and Dr Xuhua Xia (XX) is the corresponding author. This work was the result of a collaborative project between me and members of the Xia lab: RP and XX. This project was conceptualized by XX and provided supervisorial guidance, XX also proposed the indices (equation 3.1 and Figure 3.1) and models (equation 3.3 and 3.4). The development of the hypotheses, data analyses and interpretations resulted from discussions among SC, RP and XX.

3.3 Introduction

Codon adaptation has been well documented in bacterial and fungal genomes (Ikemura 1981a; Gouy, Gautier 1982; Ikemura 1992; Xia 1998) as well as in mitochondrial genomes in vertebrates (Xia 2005; Xia et al. 2007) and fungi (Carullo, Xia 2008; Xia 2008). Optimizing codon usage according to the codon usage of highly expressed host genes has been shown to increase the production of viral proteins (Haas, Park, Seed 1996; Ngumbela et al. 2008) or transgenic genes (Hernan et al. 1992; Kleber-Janke, Becker 2000; Koresawa et al. 2000). Studies on codon-anticodon adaptation have progressed in theoretical elaboration (Bulmer 1987; Bulmer 1991; Xia 1998; Higgs, Ran 2008; Jia, Higgs 2008; Xia 2008; Palidwor, Perkins, Xia 2010) as well as in critical tests of alternative theoretical predictions (Xia 1996; Xia 2005; Carullo, Xia 2008; van Weringh et al. 2011).

Codon-anticodon adaptation has been documented in phage species, in particular because several phage species have been used to treat human infections (Abedon et al. 2011) - *Staphylococcus aureus* phages (Sau et al. 2005), *Mycobacterium tuberculosis* phages

(Ranjan, Vidyarthi, Poddar 2007), *Aeromonas hydrophila* phage Aeh1 (Sau 2007) or remove infectious biofilms (Azeredo, Sutherland 2008), and need to be reengineered to improve translation efficiency. While phage codon adaptation is shaped mainly by mutation and tRNA-mediated selection, previous studies (Grosjean et al. 1978; Gouy 1987; Kunisawa, Kanaya, Kutter 1998; Sahu et al. 2005; Carbone 2008; Lucks et al. 2008) have focused mainly on the tRNA-mediated selection on codon usage. However, these studies have not assessed jointly and quantitatively the combined effect of mutation and selection on codon usage bias of phages.

Here we aim to elucidate how biased mutation and selection mediated by host translation machinery will alter the trajectory of codon adaptation in phage protein-coding genes. Many DNA phages, especially single-stranded DNA (ssDNA) phages, experience strong C→T mutation bias mediated by spontaneous or enzymatic deamination (Duncan, Miller 1980; Lindahl 1993; Xia, Yuen 2005). In particular, spontaneous deamination rate is about 100 times higher in ssDNA than in double-stranded DNA (dsDNA) based on experimental evidence (Frederico, Kunkel, Shaw 1990), which may explain why some ssDNA viruses, including ssDNA phages, evolve much faster than dsDNA viruses, with their evolutionary rate comparable to that of RNA viruses (Umemura et al. 2002; Shackelton et al. 2005; Xia, Yuen 2005; Shackelton, Holmes 2006; Duffy, Holmes 2008; Duffy, Holmes 2009).

3.3.1 The effect of C→T mutation bias

If the C→T mutation bias is strong on the phage genome, but there is no selection and no stochastic fluctuation of codon frequencies of viral protein-coding genes, then all Y-

ending codon families and subfamilies (where Y stands for pyrimidine) in viral protein-coding genes will all tend to have the same proportion of U-ending codons, i.e.,

$$P_{U.i} = \frac{N_{U.i}}{N_{U.i} + N_{C.i}} = B_{C \to T} \tag{3.1}$$

where $N_{U.i}$ and $N_{C.i}$ are the number of codons ending with U or C, respectively, in codon family *i*, and $B_{C \to T}$ is a constant representing C→T mutation bias (being 0.5 when there is no C→T mutation bias). When $B_{C \to T}$ increases, $P_{U.i}$ for all codon families will tend to increase synchronously if not checked by selection. Eq. (3.1) represents a purely mutation-only model of codon usage bias in the viral Y-ending codon families.

When the effect of selection on viral codon usage is negligible, $B_{C \to T}$ can be approximated simply as the average of $P_{U.i}$ values for all Y-ending codon families in the viral protein-coding genes, i.e.,

$$\overline{P}_U = \frac{\sum_{i=1}^{N_Y} P_{U.i}}{N_Y} \tag{3.2}$$

where N_Y is the number of codon families with Y-ending codons. For simplicity, we will refer to both Y-ending codon families (e.g., Asn codon family AAY) and subfamilies (e.g., GGY codons in Gly codon family) as Y-ending codon families.

Some ssDNA phages have high $\overline{P}_U$ values, e.g., $\overline{P}_U$ at the third codon position of *Chlamydia* phage Chp1 (Microviridae, NC_001741) is 0.9518, with U-ending codons being invariably the most frequent in all Y-ending or N-ending codon families. Some dsDNA phages can also have high $\overline{P}_U$, e.g., being 0.9014 at the third codon position of *Clostridium* phage phi3626 (Siphoviridae, NC_003524). However, codon usage bias is almost always the result of both mutation bias and selection.

3.3.2 The effect of tRNA-mediated selection and its characterization

A bacterial host may have many tRNAs to read the U-ending codons and few to read the C-ending codons in certain codon families. In such a codon family, a U-ending codon is expected to be decoded efficiently (U-friendly), i.e., tRNA-mediated selection will favor U-ending codons. Similarly, we will refer a codon in which C-ending codons can be decoded more efficiently than U-ending codons as U-hostile. A strong C→T mutation bias would accelerate/enhance codon adaptation in a U-friendly codon family, but would go against codon adaptation in U-hostile codon families. Thus, the degree of U-friendliness in the host is expected to be a major determinant of phage codon evolution.

How to measure U-friendliness (i.e., selection in favor of U-ending codons)? We develop a simple index, numerically illustrated in Figure 3.1, based on the comparison between codon frequency (CF) of all host genes and that of highly expressed host genes (CF_{AllCDS} and CF_{HEG}, respectively). Take for example the Ala (A) and Phe (F) codon families where the Y-ending codons are translated by tRNA with a wobble G. In the Ala codon family, GCC is more frequent than GCU when all coding sequences (CDSs) are included. This alone may suggest that the host translation machinery favors C-ending codons. However, in highly expressed genes, GCC is much less frequent than GCU, suggesting that U-ending codons are more efficiently translated than C-ending codons in the Ala codon family. Similarly, the Phe codon family has more UUU codons than UUC codons when all CDSs are included, but fewer UUU than UUC codons when only highly expressed genes are included. The observation that UUC is preferred by highly expressed genes suggests that the Phe codon family is not U-friendly. These illustrations lead us to adopt the association coefficient φ as a proxy for U-friendliness. The Ala codon family is U-friendly and has a

positive φ value, whereas the Phe codon family is U-hostile and has a negative φ value (Figure 3.1). φ takes values between -1 and 1 and is equivalent to the Pearson correlation coefficient for continuous variables. Because φ_i measures the selection (preference of the host machinery) in favor of the U-ending codons, it is expected to be positively correlated with $P_{U.i}$.

AA	Codon	CF_{All}	CF_{HEG}	Phi
A	GCC	34769	1306	0.1283
A	GCU	20814	2288	
F	UUC	22561	2229	-0.1348
F	UUU	30428	872	
	NNC	n_{11}	n_{12}	$n_{1.}$
	NNU	n_{21}	n_{22}	$n_{2.}$
		$n_{.1}$	$n_{.2}$	n

$$\varphi = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}}$$

Figure 3.1 - Rationale of using the phi (φ) coefficient as a proxy for U-friendliness, based on the codon frequencies (CF) between highly expressed genes (HEG) and all genes (All). φ can take values within the range between -1 and 1.

Should we develop an index of selection based only on the highly expressed genes? The following scenario suggests that we should not. Suppose the codon frequencies of NNC and NNU for the highly expressed genes is 80 and 90, respectively, but those for all CDSs are 10 and 900, respectively. A proper interpretation of this scenario is that extremely high U-biased mutation leads to the dominance of NNU codons. However, the host translation machinery prefers C-ending codons and this selection balances the U-biased mutation so that codon usage in the highly expressed genes is not as U-biased in that in all CDSs. If we have codon

usage of only highly expressed genes, we may conclude that the host translation machinery prefers U-ending codons.

3.3.3 A simple model of the joint effect of mutation and selection

The development of the φ coefficient as a measure of selection for each Y-ending codon family allows us to extend the mutation-only model in Eq. (3.1) to include the effect of selection on $P_{U.i}$, i.e.,

$$P_{U.i} = B_{C \to T} + b\varphi_i \tag{3.3}$$

Because $P_{U.i}$ can be readily computed from viral protein-coding genes, φ_i can be derived from codon frequencies of all host CDSs and of host highly expressed genes (Figure 3.1), we can use Eq. (3.3) to quantify the relative importance of mutation ($B_{C \to T}$) and selection (φ) on the codon usage bias of Y-ending codon families. If φ_i values differ little among Y-ending codon families in a host, then $P_{U.i}$ will largely depend on $B_{C \to T}$, and we will observe little variation in $P_{U.i}$ values among different codons. In contrast, with increasing intensity of selection (a large b) or increasing variation among different Y-ending codon families (i.e., large variation in φ), $P_{U.i}$ will become more dependent on $b\varphi_i$. Similarly, if $B_{C \to T}$ become very large (i.e., very strong mutation bias), then $b\varphi_i$ naturally would become relatively small and we would conclude that the mutation bias is the dominant factor in shaping codon usage in Y-ending codon families.

One may also argue that $P_{U.i}$ cannot be greater than 1 or smaller than 0, so it will asymptotically approach 1 with increasing φ_i, and approach 0 with decreasing φ_i. This implies a sigmoidal relationship between *NLM* P_U and φ. For this reason, we have also fitted the following sigmoid function:

$$NLM\ P_{U.i} = \frac{1}{1 + Ce^{-D\varphi}} \tag{3.4}$$

where parameters C and D are constants. The maximum and minimum values for $NLM\ P_U$, according to Eq.(3.4) is 1 and 0, respectively. When $\varphi = 0$ or $D = 0$, the expected $NLM\ P_U$ is $1/(1+C)$ which is equivalent to $B_{C\rightarrow T}$ in Eq.(3.3). In most cases, $B_{C\rightarrow T}$ and $1/(1+C)$ are nearly identical and we will use $B_{C\rightarrow T}$ to refer to both as an index of C→T mutation bias.

Note that for a given viral species, $B_{C\rightarrow T}$ is constant and affects uniformly the codon usage bias of all Y-ending codon families. In contrast, φ_i is specific to individual codon families. $B_{C\rightarrow T}$ is estimated by the intercept of the linear regression model and selection intensity b is the slope. Also note that the correlation coefficient between P_U and φ is also a measure of the effect of selection on codon usage bias (a measure of adaptation) in the Y-ending codon families. We interpret adaptation broadly. For example, suppose that a phage species has evolved good codon adaptation to host species A. If the phage subsequently invaded host species B, and if the codon preference in host species B is exactly the same as that in host species A, then we will state that the phage exhibit good codon adaptation to host species B, although it is pre-adaptation that is applicable here.

In this paper we use Eq. (3.3) to characterize the joint effect of mutation bias and selection by using existing genomic data from dsDNA and ssDNA phages and their hosts. We detected the effect of φ in most dsDNA and ssDNA phage species. However, increasing C→T mutation bias significantly reduced the effect of selection in ssDNA phages and shifted the phage codon usage away from the optimum. Strand asymmetry with the associated local variation in mutation bias (U-biased in some half of the genome and C-biased in the other half) can significantly interfere with codon adaptation in both dsDNA and ssDNA phages.

Much of the variation in codon adaptation among dsDNA phages can be attributed to lineage effects, with some phage lineages having uniformly strong codon adaptation and some other lineages having uniformly weak codon adaptation.

3.4 Materials and methods

3.4.1 Genomic data and processing

The genome sequences of 469 dsDNA phages, 41 ssDNA phages and their corresponding bacterial hosts were downloaded from GenBank, of which 71 have *E. coli* specified as their host in the "/HOST" tag in "FEATURES" table, including 60 dsDNA phages and 11 ssDNA phages. The coding sequences (CDSs), codon usage data, as well as three codon positions, were extracted by using DAMBE (Xia 2013b). All phage genomes were searched for encoded tRNAs by using tRNAscan-SE Search Server (Schattner, Brooks, Lowe 2005). The local TC skew plot, with the TC skew computed as $(N_T\text{-}N_C)/(N_T+N_C)$ where N_i is the number of nucleotide i along a moving window, is generated from DAMBE (Xia 2013b). All statistical analyses were done with SAS (SAS Institute 1994), with the linear regression fitted by the GLM procedure and sigmoid function by the NLIN procedure.

E. coli has 29 strains with RefSeq genomic sequences, but the "/HOST" tag in a viral genome gives only species name (i.e., *E. coli*), with no strain-specific information. For this reason, all 29 RefSeq genomic sequences were downloaded and *E. coli* codon usage is computed as the average of all CDSs from these 29 genomes. The codon usage of highly expressed *E. coli* genes was compiled in the Eeco_h.cut file distributed with EMBOSS (Rice, Longden, Bleasby 2000). It is almost perfectly correlated with our own compilation of codon usage from all *E. coli* ribosomal proteins (which are necessarily highly expressed because of

the high density of ribosomes in the cell). There is little variation in highly expressed genes among different *E. coli* genomes.

3.4.2 Indices of codon usage bias

While we mainly focus on modeling mutation and selection on P_U in Eqs. (3.3) and (3.4), two indices of codon usage bias were used to aid the interpretation of the results: codon adaptation index (CAI, Sharp, Li 1987) with the improved implementation (Xia 2007), and effective number of codons (N_c, Wright 1990) with the improved implementation (Sun, Yang, X. 2013). All these indices were computed by using DAMBE (Xia 2013b). For computing phage CAI, the host highly expressed genes is used as the reference set of genes. Only CDSs with at least 33 codons (99 nt) are included in computing the indices of codon usage bias to alleviate stochastic noise in computing these indices with few codons.

3.4.3 Phylogenetic analysis

Coancestry of phage species is difficult to establish. Although some dsDNA phage genomes are annotated to contain a DNA polymerase gene, the gene sequences from different phage lineages are often not homologous and cannot be aligned. We build phage "phylogenetic" trees by using a composition vector approach called CVTree (Xu, Hao 2009) which does not require aligned sequences. The method uses amino acid sequences, and is conceptually based on the sharing of ancient peptides that give individual evolutionary lineages its uniqueness. Computationally, the method is built upon the similarities in the sharing of words of length k ($a_1a_2\ldots a_k$) after subtracting its random expectation based on the frequencies of $a_1a_2\ldots a_{k-1}$, $a_2a_3\ldots a_k$ and $a_2a_3\ldots a_{k-1}$. The CVTree method has been implemented in the most recent version of DAMBE (Xia 2013b) because the CVTree web server does not work and the download link for the standalone program is broken. We used a

k value of 5 which has been recommended for viral genomes (Xu, Hao 2009). The data for reconstructing phylogenetic trees with the CVTree method are .faa files downloaded from GenBank, with each .faa file containing all annotated amino acid sequences for each phage species.

3.5 Results and discussion

3.5.1 Codon preference by the E. coli translation machinery: φ

tRNA-mediated selection by the host translation machinery is reflected by the codon usage of the host genes. Take the Ala codon family from *E. coli* K12 genome for example. The frequencies of GCC and GCU are 34769 and 20814 respectively, which might mislead us to think that *E. coli* translation machinery prefers codon GCC over codon GCU. However, the frequencies of GCC and GCU are 1306 and 2288, respectively, from highly expressed *E. coli* genes. This suggests that the preferred codon is GCU instead of GCC. It is for this reason that we proposed to use φ (Figure 3.1) as an index to measure the preference of U-ending codons over C-ending codons by the host translation machinery (i.e., U-friendliness).

The φ values for *E. coli* Y-ending codon families are shown in Table 3.1. Among the 16 Y-ending codon families and subfamilies, seven are U-friendly (with $\varphi > 0$, with the mean = 0.0778) and nine are U-hostile (with $\varphi < 0$, with mean = - 0.0939). Thus, C-ending codons overall should be slightly favored over U-ending codons to achieve the codon usage pattern of highly expressed host genes, which is consistent with the proportion of U-ending codons in Y-ending codon families in *E. coli* highly expressed genes ($P_{U.Ecoli} = 0.4421$). Increased C→T mutation bias will improve codon adaptation for the seven U-friendly codon families, but will lead to deterioration in the nine U-hostile codon families (Table 3.1).

Table 3.1 - Codon frequencies (CF) for Y-ending codons in *E. coli,* compiled for all coding sequences (AllCDSs) and for highly expressed genes (HEG), together with the gene copy number of tRNA in the genome (strain K12) whose anticodon matches the codon, and φ as a measure of codon preference of the host translation machinery (a large φ correspond to greater preference of U-ending codons over C-ending codons.

AA	Codon	CF_{AllCDS}	CF_{HEG}	tRNA	φ
A	**GCC**	**34769**	**1306**	**2**	**0.1283**
A	**GCT**	**20814**	**2288**		
C	TGC	8872	475	1	-0.0338
C	TGT	7072	270		
D	GAC	26012	2786	3	-0.0885
D	GAT	43817	2345		
F	TTC	22561	2229	2	-0.1348
F	TTT	30428	872		
G	**GGC**	**40405**	**2987**	**4**	**0.0495**
G	**GGT**	**33737**	**3583**		
H	CAC	13304	1160	1	-0.1226
H	CAT	17647	477		
I	ATC	34275	3488	3	-0.1111
I	ATT	41428	1640		
L	CTC	15132	541	1	-0.0338
L	CTT	15036	357		
N	AAC	29506	2832	4	-0.1383
N	AAT	24191	539		
P	**CCC**	**7481**	**38**	**1**	**0.1**
P	**CCT**	**9578**	**343**		
R	**CGC**	**30003**	**1530**		**0.0901**
R	**CGT**	**28523**	**2995**	**3**	
S	AGC	21883	1015	1	-0.0803
S	AGT	11970	168		
S	**TCC**	**11759**	**1110**	**2**	**0.0281**
S	**TCT**	**11537**	**1320**		
T	**ACC**	**31868**	**2533**	**2**	**0.0357**
T	**ACT**	**12236**	**1286**		
V	**GTC**	**20796**	**824**	**2**	**0.113**
V	**GTT**	**24966**	**2669**		
Y	TAC	16663	1569	3	-0.1018
Y	TAT	22072	865		

3.5.2 Effect of mutation and selection on codon usage of E. coli ssDNA phages

If selection in favor of U-ending codons by the host translation machinery (φ) is efficient, then we expect P_U to increase with φ. This expectation is consistent with data from *E. coli* Enterobacteria phage G4 (NC_001420) showing P_U increasing roughly linearly with φ (Figure 3.2). Fitting the linear model in Eq. (3.3) results in $B_{C\to T} = 0.5711$, and $b = 0.7242$, with the relationship being statistically significant ($p = 0.0162$). Fitting the sigmoid function in Eq. (3.4) yields $C = 0.7390$, $D = 3.0644$, and $1/(1+C) = 0.5750$ which is equivalent to $B_{C\to T}$ in the linear model, i.e., both being the expected P_U value when there is no selection. The predicted values from the linear model in Eq. (3.3) and the nonlinear model in Eq. (3.4) are identical to the first two digits after the decimal point, indicating sufficiency of the linear model.

The estimated $B_{C\to T}$ from applying the regression model in Eq. (3.3) to all 11 ssDNA Enterobacteria phages parasitizing *E. coli* varies from 0.5447 to 0.7424 (Table 3.2). These would be the P_U value when selection mediated by the host translation machinery is absent. The φ values from *E. coli* are generally small, ranging from -0.1383 to 0.1283 (Table 3.1). With the slopes in Table 3.2, the effect of selection on viral P_U values is small relative to $B_{C\to T}$. We thus expect the estimated $B_{C\to T}$ values to be close to the empirical $\overline{P}_U$ values defined in Eq.(3.2), which is true (Figure 3.3).

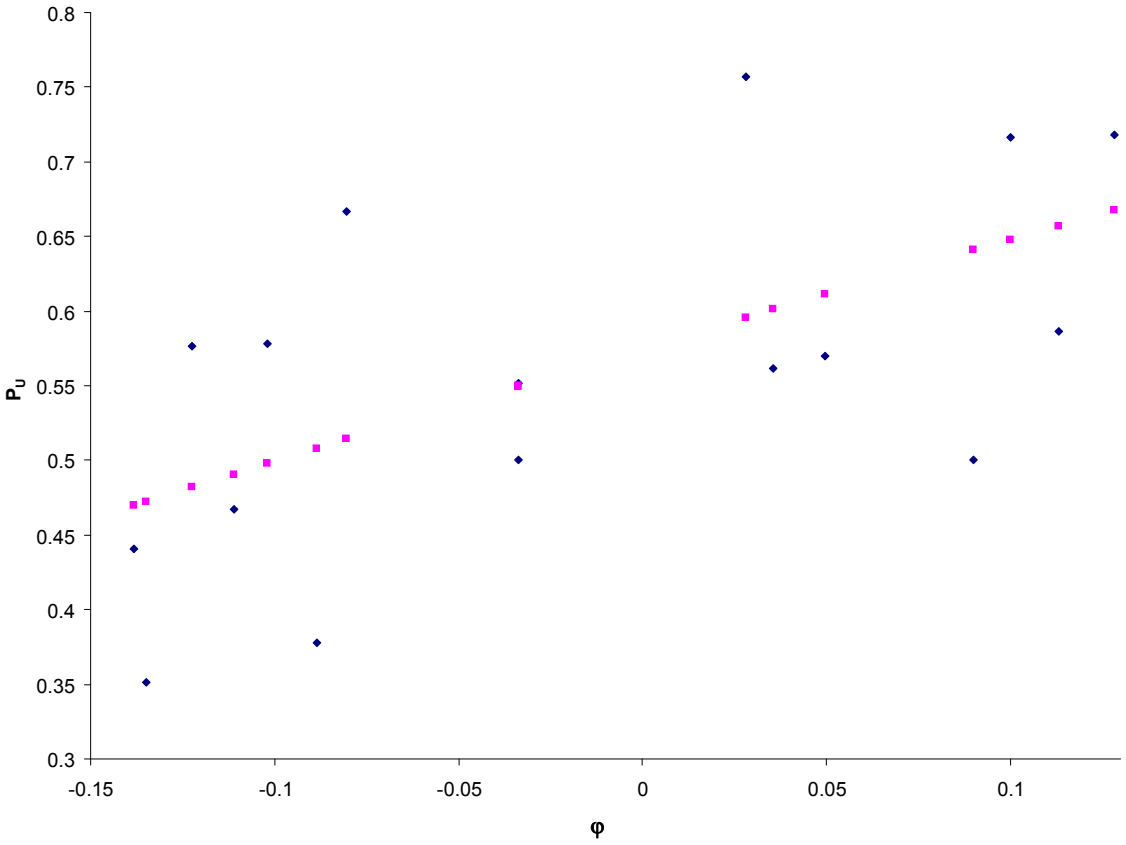

Figure 3.2 - Relationship between P_U (the proportion of U-ending codons in Y-ending codon families) and φ (selection in favor of U-ending codons), based on codon usage data from *E. coli* Enterobacteria phage G4 (NC_001420). The pink dots are the predicted values based on the sigmoid function in Eq. (3.4) and fall almost perfectly on a straight line. Applying the linear regression model in Eq. (3.3) will generate effectively the same predicted values.

Table 3.2 - Results of fitting the linear regression model in Eq. (3.3) to codon usage in ssDNA Enterobacteria phages parasitizing *E. coli,* with viral genome accession number (ACCN), viral genome length (L), number of viral genes (N_g), the estimated intercept ($B_{C\to T}$) and slope (b), the Pearson correlation between P_U and φ for each phage species, and the statistical significance (two-tailed p) of the relationship.

Phage	ACCN	L	N_g	$B_{C\to T}$	b	R	p
phage alpha3	NC_001330	6087	10	0.7026	0.4255	0.4023	0.1224
phage G4	NC_001420	5577	11	0.5711	0.7242	0.5898	0.0162
phage ID18	NC_007856	5486	11	0.5882	0.7303	0.5182	0.0398
phage ID2	NC_007817	5486	11	0.5447	0.64	0.5061	0.0455
phage phiX174	NC_001422	5386	11	0.6968	0.3171	0.3018	0.256
phage St-1	NC_012868	6094	11	0.6884	0.3251	0.3117	0.24
phage WA13	NC_007821	6068	10	0.7125	0.2478	0.2426	0.3654
phage I2-2	NC_001332	6744	9	0.6993	0.3804	0.278	0.2972
phage If1	NC_001954	8454	10	0.6552	-0.0716	-0.07	0.7966
phage Ike	NC_002014	6883	10	0.7424	0.1972	0.1221	0.6523
phage M13	NC_003287	6407	10	0.7394	0.1994	0.2008	0.4558

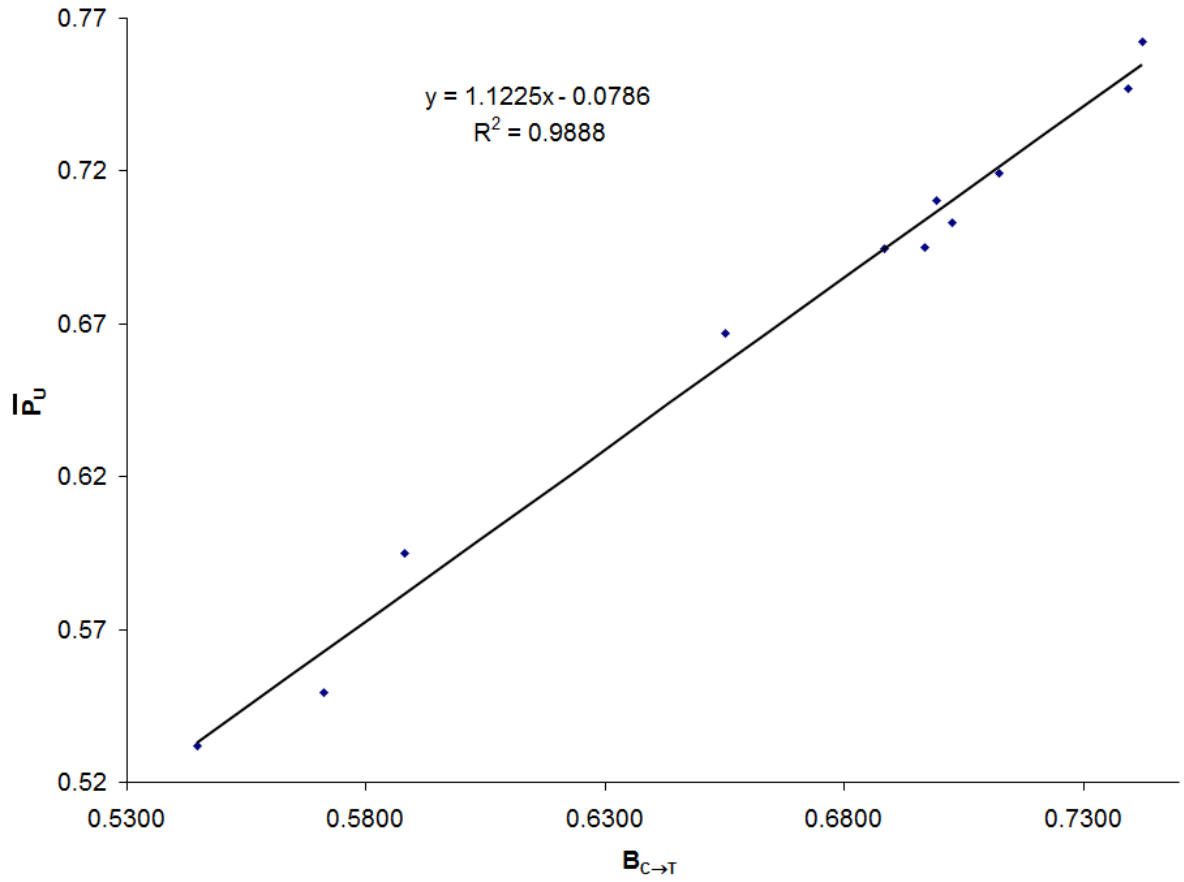

Figure 3.3 - The average $\overline{P}_U$ defined in Eq. (3.2) is similar to $B_{C\to T}$ estimated from fitting the linear model in Eq. (3.3), based on 11 ssDNA Enterobacteria phages parasitizing *E. coli* (Table 3.2).

The standard error associated with the $B_{C\to T}$ values is in the order of 0.02 (not shown), so that $B_{C\to T}$ values in Table 3.2 are all significantly greater than the observed $\overline{P}_U$ (= 0.4421) in *E. coli* highly expressed genes. If we assume that the codon usage of highly expressed *E. coli* genes represents the optimum achievable given the counterbalance between mutation and selection, then the large $B_{C\to T}$ values suggest that C→T biased mutation in ssDNA has shifted the codon usage of ssDNA phages away from the optimum.

$B_{C\to T}$ has a strong effect on effective number of codons (N_c) as expected. N_c is at its maximum when $B_{C\to T}$ is around 0.5, but decreases sharply as $B_{C\to T}$ increases leading to U-

ending codons dominating over C-ending codons. However, $B_{C\rightarrow T}$ has little effect on CAI, partly because *E. coli* translation machinery favors U-ending codons in about half of the Y-ending codon families and C-ending codons in the other half (Table 3.1). A large $B_{C\rightarrow T}$ will increase the frequency of U-ending codons in both the U-friendly and U-hostile codon families. The positive effect in the U-friendly codon families is offset by the negative effect on U-hostile codon families.

A well-adapted codon usage in a phage species in *E. coli* should have P_U positively and highly correlated with φ, i.e., large P_U in U-friendly codon families (with large φ values) and small P_U in U-hostile codon families (with small φ values). However, a strong C→T mutation bias (a large $B_{C\rightarrow T}$) will lead to high P_U in all Y-ending codon families, resulting in reduced correlation between P_U and φ. We therefore expect the correlation between P_U and φ (R) to decrease with increasing $B_{C\rightarrow T}$. This expectation is consistent with the empirical data (Figure 3.4), although there is one outlying point (Enterobacteria phage If1; NC_001954) for which we will offer an explanation later.

The four ssDNA phage species (Phage I2-2, IF1, Ike and M13) with low correlation between P_U and φ (red dots in Figure 3.4) have codon usage significantly correlated with each other, which suggests that they might be phylogenetically related. The tree built with the CVTree algorithm (Xu, Hao 2009) implemented in DAMBE (Xia 2013b) does cluster these four species into a monophyletic taxon (Figure 3.5). Other viral proteomic trees (Rohwer, Edwards 2002; Edwards, Rohwer 2005) also group these four *E. coli* ssDNA phages into the same clade.

Because of the phylogenetic structure (Figure 3.5), one may argue that the 11 points are not statistically independent and question the validity of using the conventional regression to

test the significance of the negative association between R and $B_{C\rightarrow T}$. By using the sub tree for the 11 species in Figure 3.5, we performed independent-contrasts analysis implemented in DAMBE (Xia 2013b; Xia 2013a, pp. 24-29) and found the negative association to be still significant ($p < 0.05$).

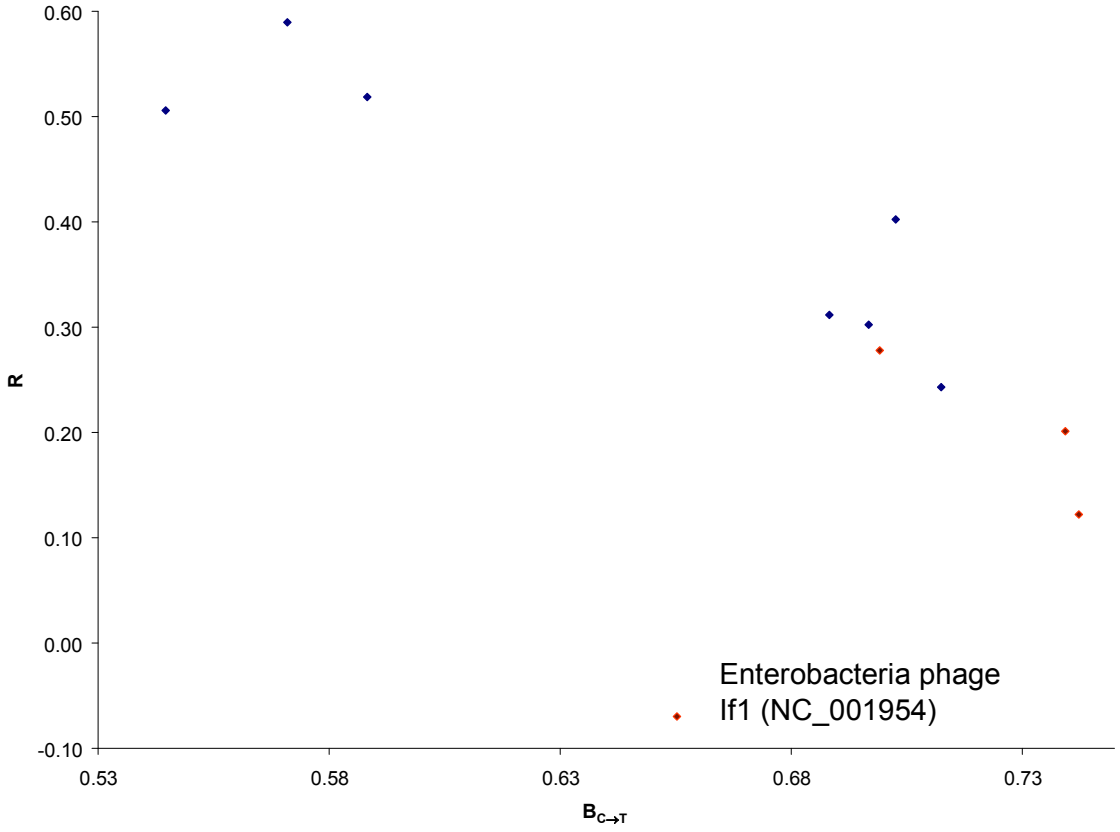

Figure 3.4 - The correlation (R) between P_U and φ decreases with increasing $B_{C\rightarrow T}$. The outline point is *E. coli* Enterobacteria phage If1 (NC_001954). The negative association is statistically significant ($p = 0.0292$ with the outlying point included). The four red dots form a monophyletic taxon and the rest form another monophyletic taxon (Figure 3.5).

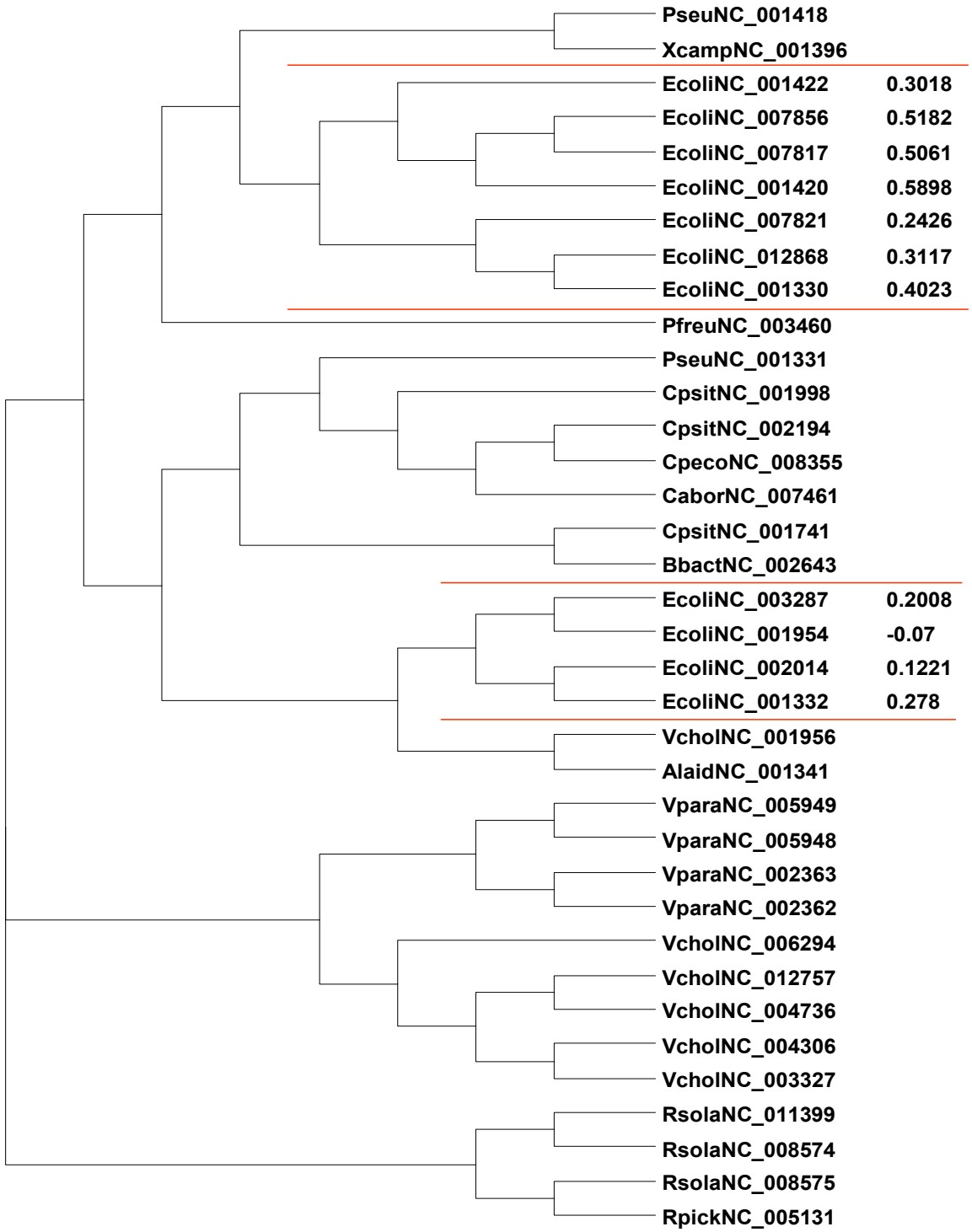

Figure 3.5 - Phylogenetic tree of ssDNA phages reconstructed by using the CVTree method (Xu, Hao 2009) implemented in (Xia 2013b). The OTUs are formed by a combination of host (the first letter of the genus name and the first four letters of the host specie name), GenBank accession number, and R (correlation between P_U and φ).

3.5.3 Effect of mutation and selection, as well as evolutionary history, on codon usage of E. coli dsDNA phages

Some dsDNA phages show strong response to selection by the host translation machinery (φ), e.g., NC_010324 phage Phieco32, with P_U strongly dependent on φ (Figure 3.6). The estimated $B_{C\rightarrow T}$ spans a wide range (Table3.3), but on average is significantly smaller than that of ssDNA phages (t-test, t = 2.1326, DF = 69, two-tailed p = 0.0365). The *b* values also vary substantially (Table3.3). For example, phage BP-4795 (NC_004813), phage cdtI (NC_009514) have negative correlations between P_U and φ. One plausible scenario for such a negative correlation to occur is when the phages have adapted to a host with codon usage quite different from that of *E. coli*, and are recent invaders of *E. coli* (i.e., they do not yet have enough time to respond to the selection of the *E. coli* translation machinery).

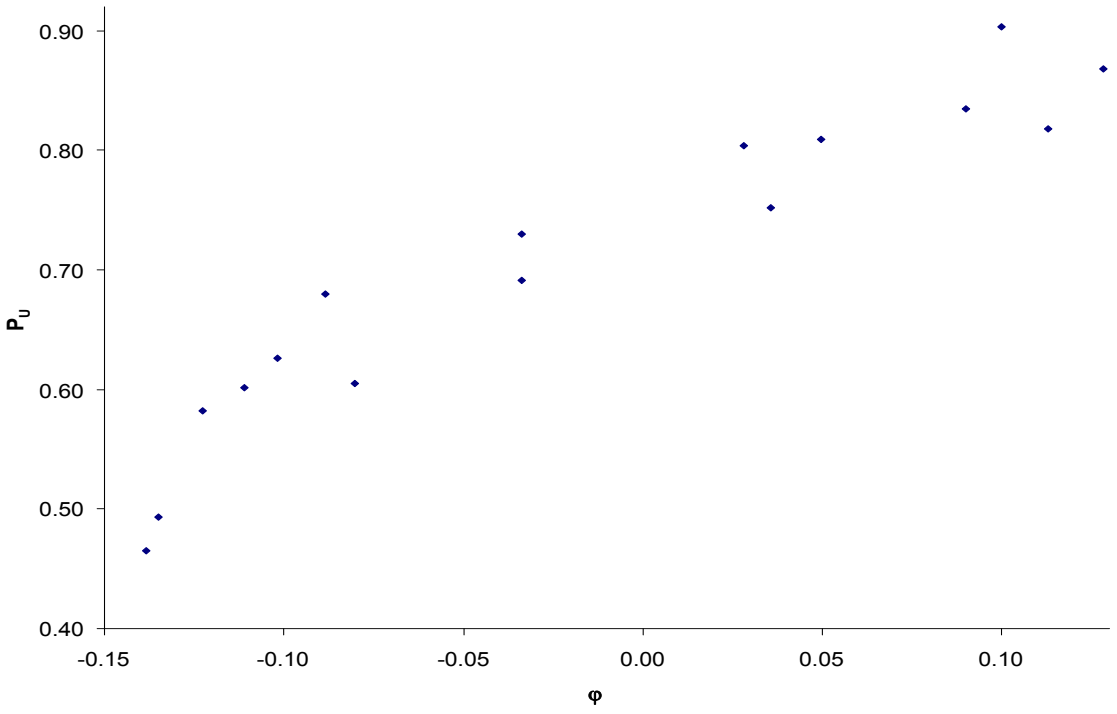

Figure 3.6 - Relationship between P_U (the proportion of U-ending codons in Y-ending codon families) and φ (selection in favor of U-ending codons), based on codon usage data from *E. coli* Enterobacteria phage Phieco32 (NC_010324).

Table 3.3 - Results of fitting the linear regression model in Eq. (3.3) to codon usage in dsDNA *E. coli* phages, with viral genome accession number (ACCN), the estimated intercept ($B_{C\to T}$) and slope (b), the Pearson correlation between P_U and φ for each phage species, and the statistical significance (two-tailed p) of the relationship.

Phage	ACCN	$B_{C\to T}$	b	R	p
phage 13a	NC_011045	0.5701	1.5227	0.8431	0
phage 285P	NC_015249	0.5781	1.6278	0.8801	0
phage 933W	NC_000924	0.5544	0.0161	0.0214	0.9373
phage BP-4795	NC_004813	0.521	-0.2274	-0.2686	0.3144
phage CC31	NC_014662	0.7285	1.1897	0.8478	0
phage cdtI	NC_009514	0.5632	-0.247	-0.287	0.2811
phage EcoDS1	NC_011042	0.5489	1.6503	0.8132	0.0001
phage EPS7	NC_010583	0.7375	0.9861	0.9041	0
phage HK022	NC_002166	0.4861	0.1869	0.1761	0.5142
phage HK97	NC_002167	0.4954	0.3254	0.2792	0.295
phage IME08	NC_014260	0.7161	0.6686	0.6667	0.0048
phage JK06	NC_007291	0.6164	0.8691	0.5367	0.0321
phage JS10	NC_012741	0.7212	0.6792	0.6625	0.0052
phage JS98	NC_010105	0.7206	0.7007	0.677	0.004
phage JSE	NC_012740	0.6937	0.6857	0.6093	0.0122
phage K1-5	NC_008152	0.6621	1.0909	0.8191	0.0001
phage K1E	NC_007637	0.6612	1.0386	0.8278	0.0001
phage K1F	NC_007456	0.6612	1.0386	0.8278	0.0001
phage lambda	NC_001416	0.4971	-0.206	-0.1824	0.499
phage Min27	NC_010237	0.5506	0.1003	0.1387	0.6086
phage Mu	NC_000929	0.4977	-0.683	-0.52	0.039
phage N15	NC_001901	0.4837	0.0368	0.0373	0.8908
phage N4	NC_008720	0.7563	1.002	0.8567	0
phage P1	NC_005856	0.5478	0.0871	0.1005	0.7111
phage P2	NC_001895	0.4767	-0.5071	-0.4678	0.0676
phage P4	NC_001609	0.5061	-0.7216	-0.5283	0.0354
phage Phi1	NC_009821	0.6911	0.6506	0.5835	0.0177
phage Phieco32	NC_010324	0.7285	1.3104	0.947	0
phage phiEcoM-GJ1	NC_010106	0.6429	1.1481	0.8759	0

Continued on next page

Continued from previous page

Phage	ACCN	$B_{C\rightarrow T}$	b	R	p
phage phiP27	NC_003356	0.545	-0.3611	-0.3513	0.1822
phage PRD1	NC_001421	0.5324	-0.6016	-0.2568	0.3371
phage RB16	NC_014467	0.6265	1.1398	0.7776	0.0004
phage RB49	NC_005066	0.6924	0.6526	0.5895	0.0163
phage RB69	NC_004928	0.7554	0.7093	0.7485	0.0009
phage RTP	NC_007603	0.61	0.979	0.5515	0.0268
phage SfV	NC_003444	0.5316	-0.1248	-0.1346	0.6193
phage SPC35	NC_015269	0.7549	0.6844	0.8722	0
phage SSL-2009a	NC_012223	0.3738	0.29	0.3104	0.242
phage T1	NC_005833	0.5795	1.0214	0.5814	0.0182
phage T3	NC_003298	0.5347	1.5114	0.8727	0
phage T4	NC_000866	0.7971	0.4221	0.5646	0.0227
phage T5	NC_005859	0.7391	0.6573	0.8449	0
phage T7	NC_001604	0.5652	1.5423	0.8505	0
phage TLS	NC_009540	0.6385	0.6612	0.3362	0.203
phage vB_EcoM-VR7	NC_014792	0.695	0.5427	0.6518	0.0062
phage VT2-Sakai	NC_000902	0.5436	0.0471	0.0588	0.8288
phage WV8	NC_012749	0.738	1.193	0.9095	0
phage bV_EcoS_AKFV3	NC_017969	0.7448	0.6505	0.8133	0.0001
phage D108	NC_013594	0.4992	-0.6469	-0.4737	0.0638
phage HK639	NC_016158	0.4496	0.2955	0.2582	0.3342
phage HK75	NC_016160	0.4842	0.3428	0.311	0.241
phage phiV10	NC_007804	0.571	0.5197	0.4588	0.0738
phage rv5	NC_011041	0.6303	0.9497	0.7012	0.0025
phage vB_EcoM_CBA12	NC_016570	0.5992	0.4134	0.4544	0.077
Stx1 converting bac	NC_004913	0.5513	0.0182	0.0227	0.9334
phage BA14	NC_011040	0.5702	1.5505	0.8803	0
Stx2 converting phage II	NC_004914	0.548	0.0595	0.0773	0.7759
Stx2-converting phage 1717	NC_011357	0.5011	-0.144	-0.1645	0.5426
Stx2-converting phage 86	NC_008464	0.5682	0.0354	0.0471	0.8625
Stx2 converting phage I	NC_003525	0.5133	-0.0575	-0.0912	0.7369

We noted that all phages that have negative correlation between P_U and φ share similar codon usages (Figure 3.7). For example, P_U values from phage BP-4795 (NC_004813) and those from phage cdtI (NC_009514) have a correlation coefficient of 0.9349. The

observation that they have similar codon usages that are different from that of their host increases the plausibility that they may have adapted to a common host that have codon usage different from that of *E. coli*.

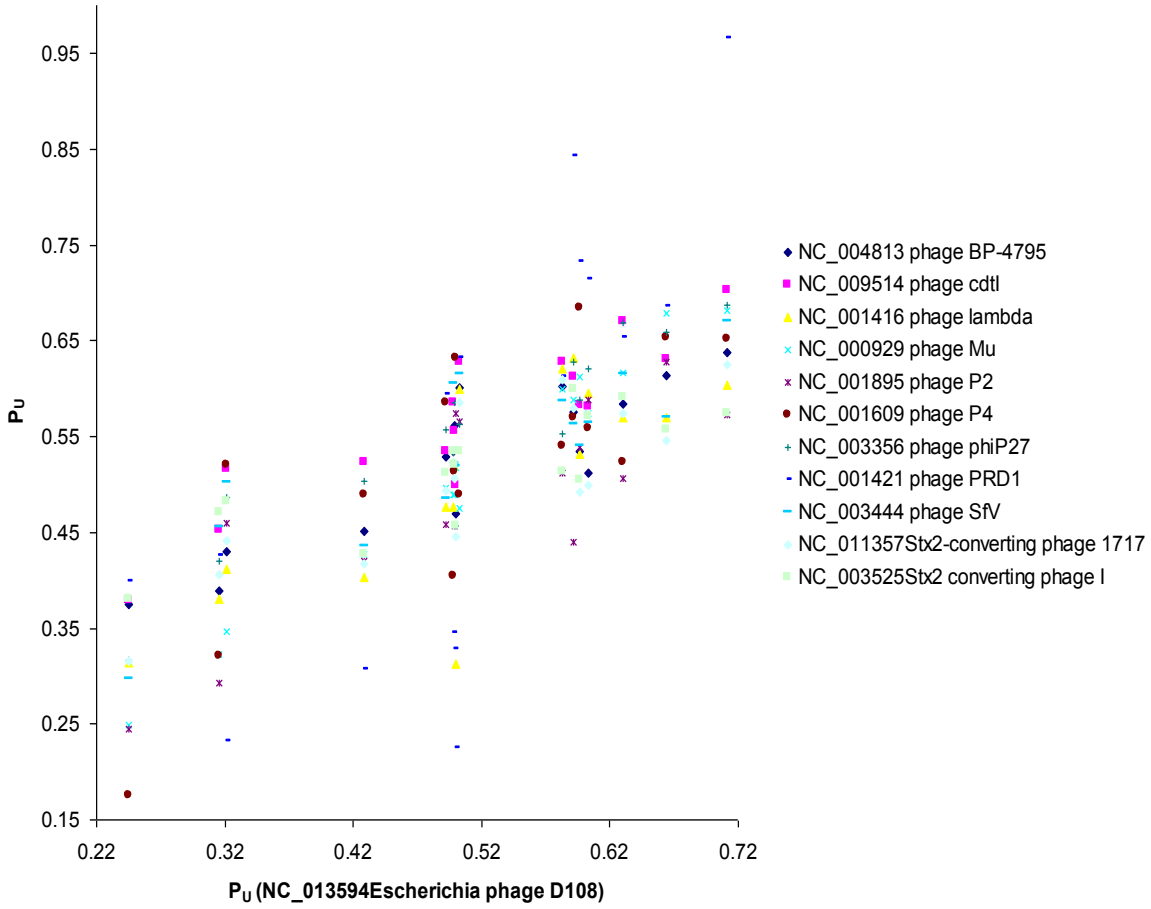

Figure 3.7 - Relationship in P_U (the proportion of U-ending codons in Y-ending codon families) among different dsDNA phage species in which P_U and φ (selection in favor of U-ending codons) are negatively correlated. The phage genomes are identified by its GenBank accession number and its species name.

Phylogenetic reconstruction with the CVTree method (Xu, Hao 2009) suggests that evolutionary history may have contributed to the differences in codon adaptation in *E. coli* dsDNA phages. Descendents of Node A (Figure 3.8) all have high R (correlation between P_U

and φ) values and a narrow range of $B_{C\rightarrow T}$ values. None of them encode tRNA genes in their genomes, whereas tRNA genes are present in many other *E. coli* dsDNA phage lineages. Their closest sister group, consisting of Enterobacteria phage WV8 (Myoviridae, NC_012749) and *Erwinia* phage phiEa21-4 (Myoviridae, NC_011811) both have many tRNA genes, with the former having 19 and the latter 23 tRNA genes. Enterobacteria phage WV8 has excellent codon adaptation, with R between P_U and φ being 0.9095. One possible scenario is that the ancestor of dsDNA phages in Cluster A (Figure 3.8) was like Enterobacteria phage WV8 which carries nearly a full set of tRNA genes and can potentially exploit a variety of bacterial host translation machinery. Upon establishing in *E. coli* which happens to have a tRNA pool favoring the codon usage of the ancestral phage, there was then no selection to maintain the set of tRNA genes, leading to the loss of all tRNA genes in dsDNA phages in Cluster A (Figure 3.8).

One may wonder why Enterobacteria phage WV8, with excellent codon adaptation at least for the Y-ending codon families and subfamilies, should still keep its set of tRNA genes. One possibility is that it is a generalist with more than one host. Previous studies have already suggested an association of host diversity and the number of tRNA genes carried on phage (Sau et al. 2007; Enav, Beja, Mandel-Gutfreund 2012). Another possibility is that the Enterobacteria phage WV8 is already in the process of losing its tRNA genes. First, it has fewer tRNA genes than its sister lineage *Erwinia* phage phiEa21-4 which has 23 tRNA genes. Second, the tRNA gene sequences in Enterobacteria phage WV8 tend to have longer branches when clustered with other tRNA genes in dsDNA *E. coli* phages, suggesting that they may be accumulating more substitutions than other presumably functional tRNA genes. This indicates weakened selection in the tRNA genes of Enterobacteria phage WV8. A

possible statistical test would be to check if mutations occur equally frequently in loop and stems region of the tRNA sequences because a functional tRNA sequences tend to accumulate more substitutions in loop regions than in stems which are crucial for maintaining the secondary structure for normal functioning. However, the tRNA sequences in these phages are all too diverged to allow accurate inference of positional substitutions.

Descendents from Nodes B and C have low R values (between P_U and φ). While those in Cluster C may indeed be recent invaders of *E. coli,* it is difficult to say the same for those under Node B because they must have been associated with *E. coli* for a long time given the diversity of phage lineages. This forces us to reexamine the assumptions of our model in Eqs. (3.3) and (3.4) in search for an alternative explanation for the lack of selection in those species in Cluster B. In the two models specified in Eqs. (3.3) and (3.4), we have assumed a uniform mutation bias that will affect all genes and all Y-ending codon families. However, because of the asymmetric ways of genome replication, the two strands often experience differential mutation bias leading to strand asymmetry documented in organisms spanning from viruses, organelles, prokaryotic and eukaryotic genomes (Marin, Xia 2008; Xia 2012b; Xia 2012a).

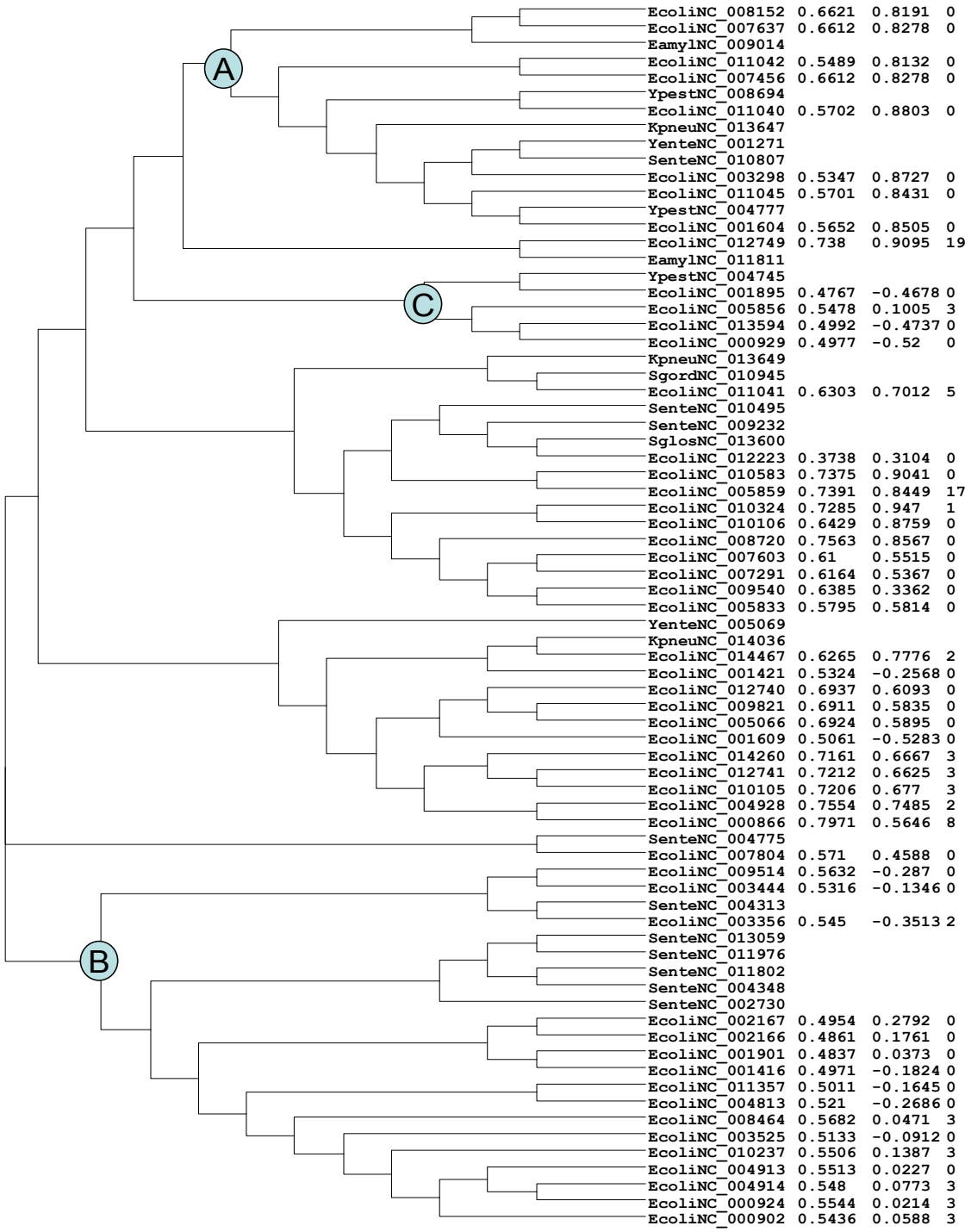

Figure 3.8 - Phylogenetic tree of dsDNA phages reconstructed by using the CVTree method with k = 5. The OTUs are formed by a combination of host (the first letter of the genus name and the first four letters of the host specie name), GenBank accession number, estimated $B_{C\to T}$, R (correlation between P_U and φ), and number of tRNA genes in the phage genome.

3.5.4 Strand asymmetry, local mutation bias and phage codon adaptation

We found that *E. coli* dsDNA phages with negative *R* values all exhibit strong local strand asymmetry similar to that of phage Mu (NC_000929) in Figure 3.9a, with the TC skew, defined as $(N_T\text{-}N_C)/ (N_T+N_C)$, spanning a wider range than those *E. coli* dsDNA phage genomes with large positive *R* values. In particular, genes with codons in the U-friendly codon families tend to be located in the relatively T-poor regions. Thus, although the *E. coli* translation machinery favors U-codons in these U-friendly codon families (a large φ), the constraint that the genes are located in the T-poor region prevents the genes to increase P_U in these codon families resulting in a low or even negative correlation between P_U and φ. In contrast, those *E. coli* dsDNA phages with high *R* values between P_U and φ exhibit relatively little local strand asymmetry as shown for phage BA14 (Figure 3.9b). Thus, strong strand asymmetry with the associated local variation in mutation bias can reduce the efficiency of selection favoring codon adaptation in dsDNA phages.

The effect of strand asymmetry on codon adaptation observed in dsDNA phages is also visible in ssDNA phages. To show this quantitatively, we used the same window size and step size and computed the variance of the window-specific skew values as an index of strand asymmetry (I_{SA}). The *R* value is highly significantly and negatively correlated with I_{SA} ($p = 0.0008$, Figure 3.10). The same negative relationship between *R* and I_{SA} holds for dsDNA phages with $p < 0.0001$. Thus, mutation bias along different parts of the phage genome in opposite directions can significantly reduce the efficiency of selection on codon usage of both ssDNA and dsDNA phages. To properly assess the effect of mutation bias and selection by the host translation machinery, it is important to apply Eqs. (3.3) and (3.4) to phage genomic segments with relatively homogenous mutation bias.

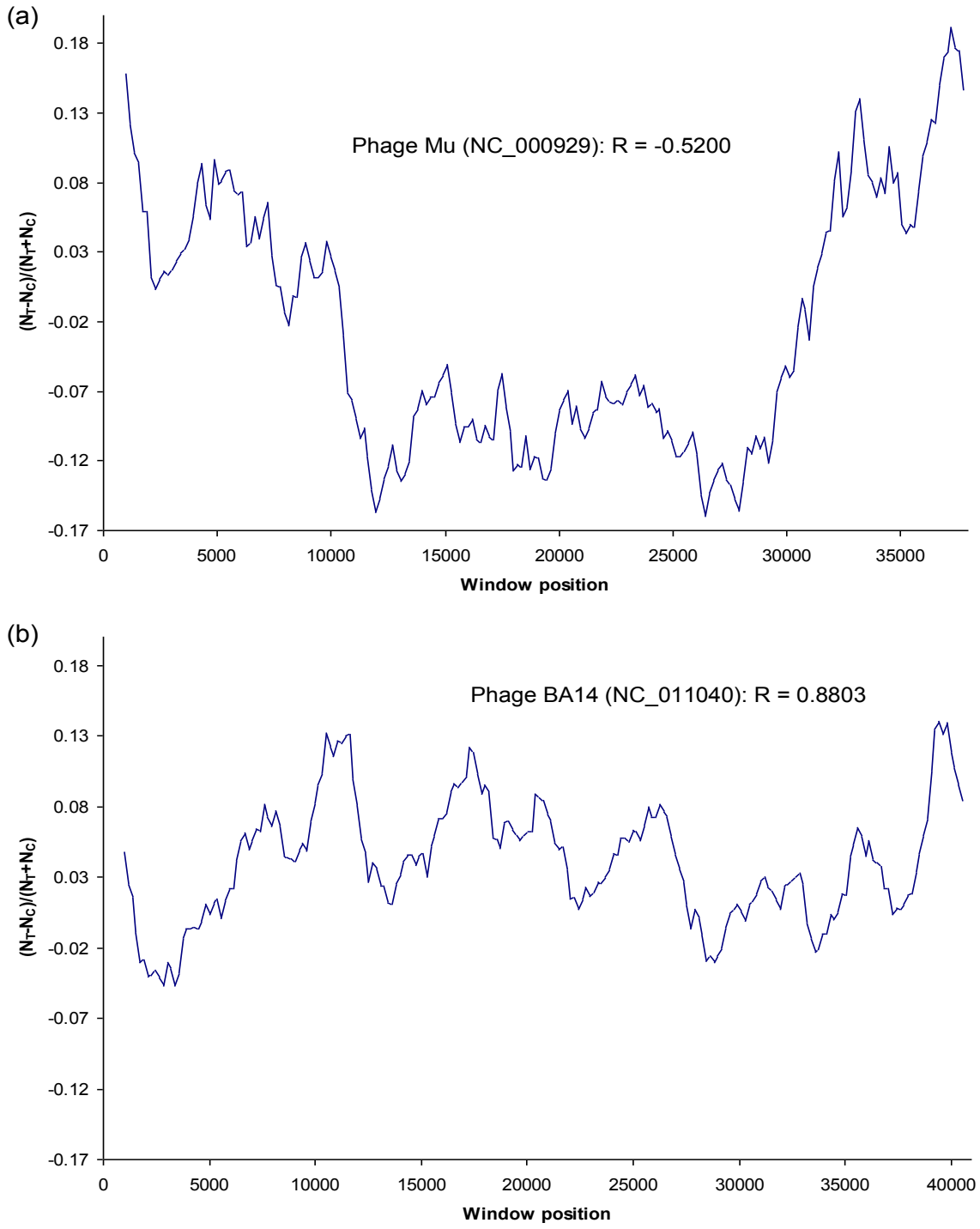

Figure 3.9 - TC skew plots for two *E. coli* dsDNA phages with different R values indicated, and with the same scale for Y-axis for visual comparison of variation in TC skew, defined as $(N_T\text{-}N_C)/ (N_T+N_C)$. Generated from DAMBE with the same windows size (= 1991 nt) and step size (= 183 nt).

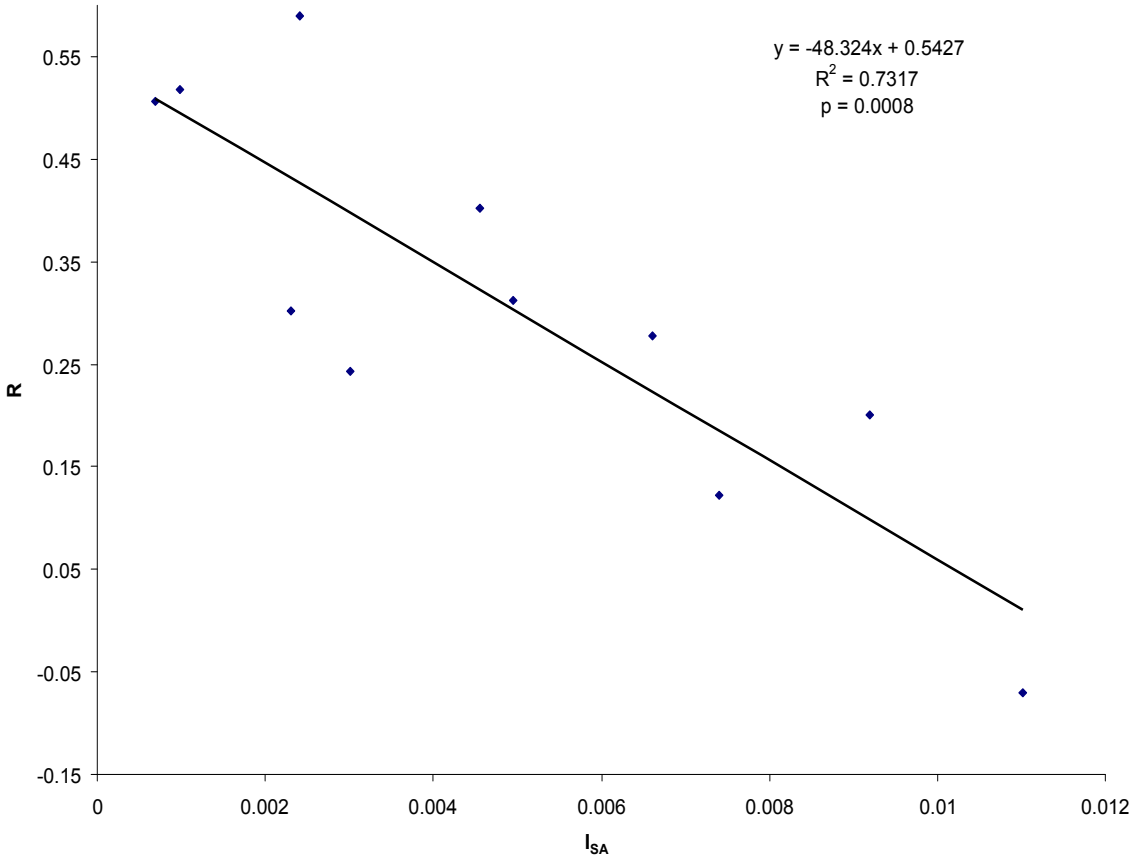

Figure 3.10 - The effect of selection on Y-ending codons, measured by the correlation (*R*) between P_U and φ, decreases with the degree of strand asymmetry, measured by the index of strand asymmetry (I_{SA}, which is the variance of the window-specific TC skew values).

The relationship between *R* and I_{SA} in Figure 3.10 offers an explanation for the outlying point in Figure 3.4, where Enterobacteria phage If1 (NC_001954) has an *R* value much smaller than expected from the general trend. This phage has the strongest strand asymmetry (i.e., the largest I_{SA} value), and in this new light is expected to be associated with a low *R* value (Figure 3.10).

3.5.5 Effect of mutation bias on nonsynonymous substitutions

The C→T mutation bias ($B_{C\to T}$) is expected not only to affect nucleotide composition at third codon positions as shown in Figure 3.3 for ssDNA phages, but also that at first and second codon positions if purifying selection is not strong enough to counteract against it. This predicts that TC skew at the second codon position should also increase with $B_{C\to T}$, a prediction consistent with the empirical evidence for both ssDNA phages (Figure 3.11a) and dsDNA phages (Figure 3.11b).

While the slope for dsDNA phages is greater than that for ssDNA phages, the difference is not statistically significant ($p = 0.8099$). This allows us to perform an ANCOVA to test the difference in intercept between the dsDNA and ssDNA phages. The resulting equations have the same slopes but significantly different intercepts ($p < 0.0001$):

$$\begin{aligned} TCSkew_{ssDNA} &= -0.2122 + 0.3986 B_{C\to T} \\ TCSkew_{dsDNA} &= -0.1455 + 0.3986 B_{C\to T} \end{aligned} \tag{3.5}$$

Thus, given the same C→T mutation bias, one expect significantly more C→T mutations at the second codon position in ssDNA phages than dsDNA phages. This may explain why genome lengths of ssDNA phages are about one order smaller than those of dsDNA phages. The inverse relationship between genome length and mutation rate have been expressed both verbally (Drake 1999) and symbolically (Xia et al. 2006) before.

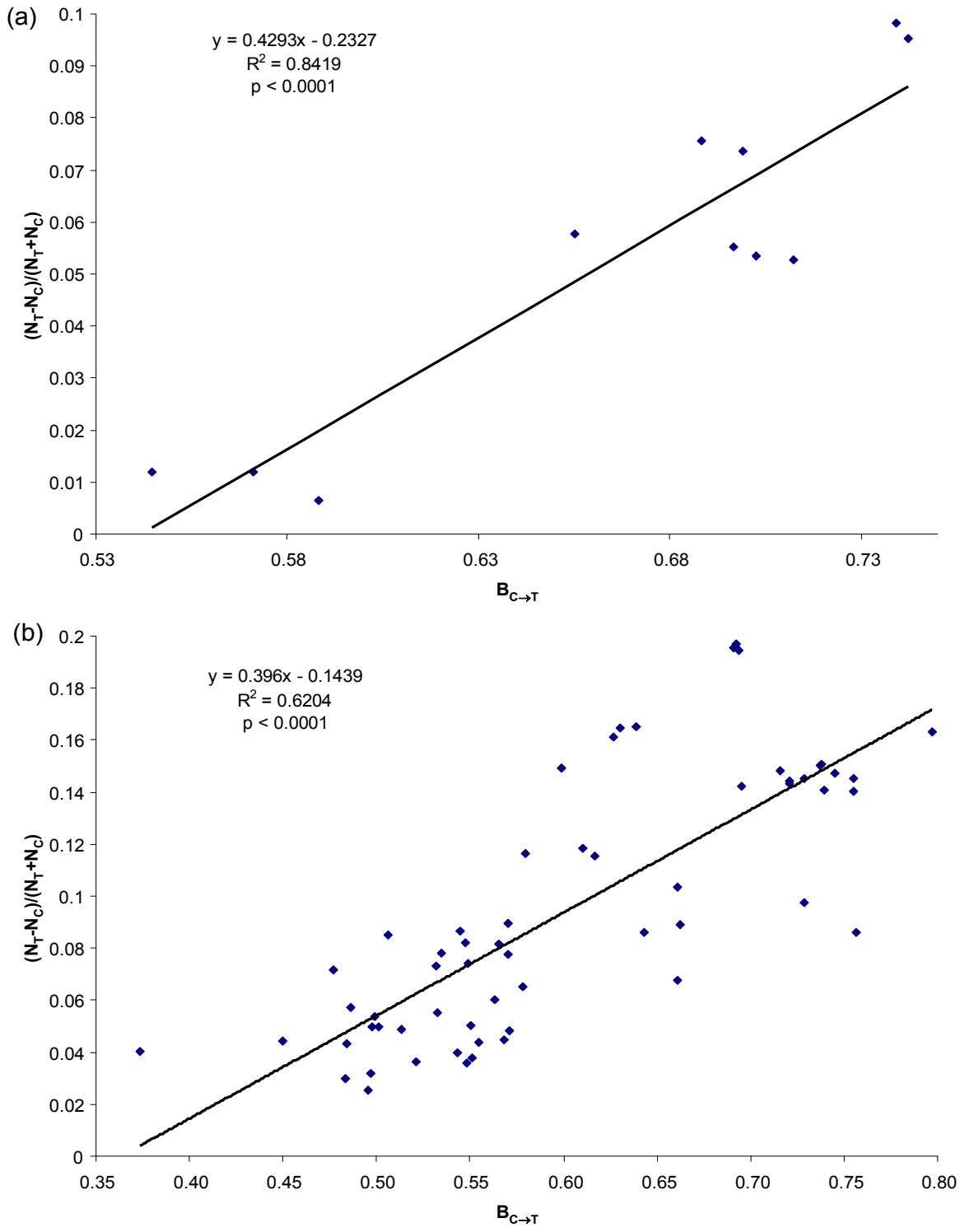

Figure 3.11 - TC skew, defined as $(N_T\text{-}N_C)/(N_T\text{+}N_C)$, at second codon position increases with C→T mutation bias ($B_{C\rightarrow T}$), suggesting the effect of mutation bias on nonsynonymous substitution. (a) ssDNA phages, (b) dsDNA phages.

3.5.6 Other factors that may contribute to phage codon usage

One factor that may contribute to phage codon usage bias is phage-encoded tRNA genes. Notice that *E. coli* dsDNA phages other than those in Clusters A, B and C (Figure 3.8) are scattered in clusters that frequently have phage lineages with phage-encoded tRNA genes. The presence of phage-encoded tRNA genes can alter the host tRNA pool, so that φ may no longer reflect the selection on codon usage. For example, if the host tRNA for an NNY codon favors U-ending codons, but phage-encoded tRNA favors C-ending codons, then φ would not be a good predictor of phage codon usage. It is noteworthy that phage species in Cluster A that do not have phage-encoded tRNA genes all have uniformly high *R* values, suggesting that selection by the host translation machinery may be more effective on phage codon usage in phages with no tRNA encoded in the phage genome. Alteration of the host tRNA pool through selective local tRNA enrichment in favor of viral gene translation has been documented in several viral species including HIV-1 (van Weringh et al. 2011) and vaccinia and influenza A (Pavon-Eternod et al. 2013).

While almost all Y-ending codons are translated by tRNAs with a wobble G (except for Ile and Arg codon families where Y-ending codons are decoded by tRNAs with a wobble A chemically modified to inosine), different tRNAs with a wobble G appear to have different codon preferences with some favoring C-ending codons, some U-ending codons, and some with no detectable preference. At present, such a preference is not well understood and cannot be properly measured. This is in contrast to R-ending codon families which are typically translated by two types of tRNAs, one with a wobble U consistently preferring A-ending codons and the other a wobble C consistently preferring G-ending codons. To properly assess the effect of phage-encoded tRNAs on phage codon usage, one needs

minimally to assess whether the tRNA is actually functional and measure the synonymous codon preference of phage tRNAs. Currently we have no means of doing this.

Although our focus is on the joint effect of mutation and host tRNA-mediated selection on phage codon usage, we are aware of other factors that have been suggested to affect codon usage. Some bacterial hosts live in high-temperature environment and have relatively high genomic GC. Their phages are also expected to have high GC to maintain genome stability in high temperature (Xia, Yuen 2005). Different host species may have the four nucleotides in quite different concentrations (e.g., nucleotide C is typically rare and A typically abundant) and cytoplasmic parasites or symbionts such as virus or organelles should avoid using rare nucleotides in building their genome and RNA molecules (Xia 1996; Xia, Palidwor 2005; Marin, Xia 2008). Such avoidance would also be reflected in codon usage bias. However, these effects are likely additive and not expected to confound the relationships we aim to study here.

3.5.7 The CVTree method for phylogenetic reconstruction

The CVTree method has been shown to be effective in reconstructing phylogenetic relationships among viruses and prokaryotic genomes based on unaligned amino acids (Xu, Hao 2009). However, its rationale or theoretical basis has never been explicitly outlined. We may imagine that the RNA world was littered with short peptides formed spontaneously at the air-water interface (Griffith, Vaida 2012). Certain peptides then form complexes with RNA to aid RNA replication. If a certain RNA molecule then encodes certain peptide that aid the replication of the RNA, then the two partners would have an obvious selective advantage over others and eventually lead to the formation of ancient “species” characterized

by self-replicating RNA-peptide complexes. Thus, the first "genome" would be an RNA molecule that may encode one or multiple copies of the peptide that aid the RNA replication.

These ancient "genomes" would then be subject to mutation, selection, recombination and horizontal transfer that gradually erodes the identity of the genome-specific peptides. Out of this entangled bank of frolicking genomes arose probably many evolutionary lineages with gradually reduced rate of horizontal gene transfer confined mainly within individual lineages (Xia, Yang 2013). Only three (Archaea, Eubacteria, and Eukarya) of these ancient lineages have representatives surviving to this day. However, the fingerprint contained in the ancient genome-specific peptides may last to this day and can be mined to trace the evolutionary history back to time immemorial. Whether this scenario captures some scientific truth remains to be tested.

4 Chapter Four

Conclusions

Phages depend on the translational machinery of their host. In order to maximize their translational efficiency, we expect the phages to maintain their codon usage in close agreement with that of host codon usage. However, on completion of a comprehensive study on the codon usage correlation between phages and their hosts we observed a wide variation in the degree of concordance in them. This observation prompted us to examine the differential factors shaping codon usage of dsDNA and ssDNA phages.

We wish to highlight our three findings in summary. First, the number of tRNA genes carried by some of the dsDNA phage genomes reduced the need for the phages to evolve a codon usage pattern similar to that of their hosts. Second, ssDNA phages, because of their high mutation rate, especially C→T mutations, prevented them from evolving a codon usage pattern as close to that of the host as dsDNA phage. Third, ssDNA phages have evolved a previously unknown type of codon adaptation by biased usage of codon families. That is, they overuse codon families in which their C→T mutations would accelerate/enhance their codon adaptation (i.e., codon families in which the host translation machinery favor U-ending codons), but reduce the use of codon families in which their C→T mutations would thwart selection for codon adaptation.

Our results show that previous studies on phage codon adaptation are insufficient in at least two ways. First, codon frequencies from either all host CDSs or all highly expressed host genes are insufficient to capture the selection by host translation machinery. Second, it is crucially important to have explicit models to dissect the effect of mutation and selection.

Our index (φ) is a proper measure of selection imposed by host translation machinery on phage codon usage, and our linear and nonlinear models allow us to estimate the C→T mutation bias ($B_{C\to T}$) and to evaluate the relative effect of the mutation bias and host translation machinery on phage codon usage. The C→T mutation bias affects not only synonymous codon usage, but also nonsynonymous substitutions at second codon positions, especially in ssDNA phages. The host translation machinery affects phage codon adaptation in both dsDNA and ssDNA phages. dsDNA phages without phage-encoded tRNA genes tend to have better codon adaptation than those with tRNA. Strand asymmetry strongly influences the efficiency of selection on codon adaptation for the simple reason that, if mutation is strongly U-biased in half of the phage genome and C-biased in the other half, selection favoring C-ending codons by host translation machinery would be efficient in only half of the phage genome.

5 References

Abedon, ST, SJ Kuhl, BG Blasdel, EM Kutter. 2011. Phage treatment of human infections. Bacteriophage 1:66-85.

Ackermann, HW. 2007. 5500 Phages examined in the electron microscope. Arch Virol 152:227-243.

Azeredo, J, IW Sutherland. 2008. The use of phages for the removal of infectious biofilms. Curr Pharm Biotechnol 9:261-266.

Bailly-Bechet, M, M Vergassola, E Rocha. 2007. Causes for the intriguing presence of tRNAs in phages. Genome Research 17:1486-1495.

Barnes, DE, T Lindahl. 2004. Repair and genetic consequences of endogenous DNA base damage in mammalian cells. Annu Rev Genet 38:445-476.

Beletskii, A, AS Bhagwat. 1996. Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in Escherichia coli. Proc Natl Acad Sci U S A 93:13919-13924.

Bennetzen, JL, BD Hall. 1982. Codon selection in yeast. J Biol Chem 257:3026-3031.

Brussow, H, E Kutter. 2005. Genomics and evolution of tailed phages. In: E Kutter, A Sulakvelidze, editors. Bacteriophages: biology and applications. Boca Raton: CRC Press. p. 91-128.

Bulmer, M. 1987. Coevolution of codon usage and transfer RNA abundance. Nature 325:728-730.

Bulmer, M. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897-907.

Burrowes, B, DR Harper, J Anderson, M McConville, MC Enright. 2011. Bacteriophage therapy: potential uses in the control of antibiotic-resistant pathogens. Expert Rev Anti Infect Ther 9:775-785.

Carbone, A. 2008. Codon bias is a major factor explaining phage evolution in translationally biased hosts. Journal of Molecular Evolution 66:210-223.

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

Carullo, M, X Xia. 2008. An Extensive Study of Mutation and Selection on the Wobble Nucleotide in tRNA Anticodons in Fungal Mitochondrial Genomes. Journal of Molecular Evolution 66:484-493.

Chattopadhyay, S, RK Ghosh. 1988. Characterization of phage-specific transfer RNA molecules coded by Vibrio eltor phage e4. Virology 165:606-608.

Clark, JR, JB March. 2004. Bacterial viruses as human vaccines? Expert Rev Vaccines 3:463-476.

Coulondre, C, JH Miller, PJ Farabaugh, W Gilbert. 1978. Molecular basis of base substitution hotspots in Escherichia coli. Nature 274:775-780.

Deschavanne, P, MS DuBow, C Regeard. 2010. The use of genomic signature distance between bacteriophages and their hosts displays evolutionary relationships and phage growth cycle determination. Virol J 7:163.

Desiere, F, C Mahanivong, AJ Hillier, PS Chandry, BE Davidson, H Brussow. 2001. Comparative genomics of lactococcal phages: insight from the complete genome sequence of Lactococcus lactis phage BK5-T. Virology 283:240-252.

Drake, JW. 1991. A constant rate of spontaneous mutation in DNA-based microbes. Proc Natl Acad Sci U S A 88:7160-7164.
Drake, JW. 1993. Rates of spontaneous mutation among RNA viruses. Proc Natl Acad Sci U S A 90:4171-4175.
Drake, JW. 1999. The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes. Annals of the New York Academy of Sciences 870:100-107.
Duffy, S, EC Holmes. 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, EC Holmes. 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, LA Shackelton, EC Holmes. 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet 9:267-276.
Duncan, BK, JH Miller. 1980. Mutagenic deamination of cytosine residues in DNA. Nature 287:560-561.
Edwards, RA, F Rohwer. 2005. Viral metagenomics. Nat Rev Microbiol 3:504-510.
Elena, SF, R Miralles, JM Cuevas, PE Turner, A Moya. 2000. The two faces of mutation: extinction and adaptation in RNA viruses. IUBMB Life 49:5-9.
Enav, H, O Beja, Y Mandel-Gutfreund. 2012. Cyanophage tRNAs may have a role in cross-infectivity of oceanic Prochlorococcus and Synechococcus hosts. ISME J 6:619-628.
Felsenstein, J. 1985. Phylogenies and the comparative method. Amer. Nat. 125: 1-15.
Fenner, F. 1971. The nomenclature and classification of viruses the International Committee on Nomenclature of Viruses. Virology 46:979-980.
Francino, MP, L Chao, MA Riley, H Ochman. 1996. Asymmetries generated by transcription-coupled repair in enterobacterial genes. Science 272:107-109.
Francino, MP, H Ochman. 1997. Strand asymmetries in DNA evolution. Trends Genet 13:240-245.
Frederico, LA, TA Kunkel, BR Shaw. 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry 29:2532-2537.
Gabbara, S, M Wyszynski, AS Bhagwat. 1994. A DNA repair process in Escherichia coli corrects U:G and T:G mismatches to C:G at sites of cytosine methylation. Mol Gen Genet 243:244-248.
Gladstone, EG, IJ Molineux, JJ Bull. 2012. Evolutionary principles and synthetic biology: avoiding a molecular tragedy of the commons with an engineered phage. J Biol Eng 6:13.
Goodridge, LD. 2010. Designing phage therapeutics. Curr Pharm Biotechnol 11:15-27.
Gouy, M. 1987. Codon contexts in enterobacterial and coliphage genes. Molecular Biology and Evolution 4:426-444.
Gouy, M, C Gautier. 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res 10:7055-7074.
Greenbaum, BD, AJ Levine, G Bhanot, R Rabadan. 2008. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathog 4:e1000079.

Griffith, EC, V Vaida. 2012. In situ observation of peptide bond formation at the water-air interface. Proceedings of the National Academy of Sciences of the United States of America 109:15697-15701.
Grosjean, H, D Sankoff, WM Jou, W Fiers, RJ Cedergren. 1978. Bacteriophage MS2 RNA: a correlation between the stability of the codon: anticodon interaction and the choice of code words. J Mol Evol 12:113-119.
Haas, J, E-C Park, B Seed. 1996. Codon usage limitation in the expression of HIV-1 envelope glycoprotein. Current Biology 6:315-324.
Hernan, RA, HL Hui, ME Andracki, RW Noble, SG Sligar, JA Walder, RY Walder. 1992. Human hemoglobin expression in Escherichia coli: importance of optimal codon usage. Biochemistry 31:8619-8628.
Higgs, PG, W Ran. 2008. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. Mol Biol Evol 25:2279-2291.
Hilterbrand, A, J Saelens, C Putonti. 2012. CBDB: the codon bias database. BMC Bioinformatics 13:62.
Hoelzer, K, LA Shackelton, CR Parrish. 2008. Presence and role of cytosine methylation in DNA viruses of animals. Nucleic Acids Res 36:2825-2837.
Hopfield, JJ. 1974. Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. Proc Natl Acad Sci U S A 71:4135-4139.
Ikemura, T. 1981a. Correlation between the abundance of *Escheriachia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. Journal of Molecular Biology 146:1-21.
Ikemura, T. 1981b. 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. 1992. Correlation between codon usage and tRNA content in microorganisms. In: DL Hatfield, BJ Lee, RM Pirtle, editors. Transfer RNA in protein synthesis. Boca Raton: CRC Press. p. 87-111.
Jia, W, PG Higgs. 2008. Codon usage in mitochondrial genomes: distinguishing context-dependent mutation from translational selection. Molecular Biology and Evolution 25:339-351.
Jukes, TH, S Osawa. 1990. The genetic code in mitochondria and chloroplasts. Experientia 46:1117-1126.
Kano-Sueoka, T, JR Lobry, N Sueoka. 1999. Intra-strand biases in bacteriophage T4 genome. Gene 238:59-64.
Karlin, S. 1999. Bacterial DNA strand compositional asymmetry. Trends Microbiol 7:305-308.
Karlin, S, W Doerfler, LR Cardon. 1994. Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? J Virol 68:2889-2897.
Kleber-Janke, T, WM Becker. 2000. Use of modified BL21(DE3) Escherichia coli cells for high-level expression of recombinant peanut allergens affected by poor codon usage. Protein Expression & Purification 19:419-424.

Koresawa, Y, S Miyagawa, M Ikawa, K Matsunami, M Yamada, R Shirakura, M Okabe. 2000. Synthesis of a new Cre recombinase gene based on optimal codon usage for mammalian systems. J Biochem 127:367-372.
Kudla, G, AW Murray, D Tollervey, JB Plotkin. 2009. Coding-Sequence Determinants of Gene Expression in Escherichia coli. Science 324:255-258.
Kunisawa, T. 1992. Synonymous codon preferences in bacteriophage T4: a distinctive use of transfer RNAs from T4 and from its host Escherichia coli. J Theor Biol 159:287-298.
Kunisawa, T. 2000. Functional role of mycobacteriophage transfer RNAs. J Theor Biol 205:167-170.
Kunisawa, T, S Kanaya, E Kutter. 1998. Comparison of synonymous codon distribution patterns of bacteriophage and host genomes. DNA Res 5:319-326.
Lim, VI. 1994. Analysis of action of wobble nucleoside modifications on codon-anticodon pairing within the ribosome. J Mol Biol 240:8-19.
Lima-Mendez, G, A Toussaint, R Leplae. 2007. Analysis of the phage sequence space: the benefit of structured information. Virology 365:241-249.
Lindahl, T. 1993. Instability and decay of the primary structure of DNA. Nature 362:709-715.
Lobo, FP, BE Mota, SD Pena, V Azevedo, AM Macedo, A Tauch, CR Machado, GR Franco. 2009. Virus-host coevolution: common patterns of nucleotide motif usage in Flaviviridae and their hosts. PLoS One 4:e6282.
Lobry, JR. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 13:660-665.
Lobry, JR. 2004. Life history traits and genome structure: aerobiosis and G+C content in bacteria. Lecture Notes in Computer Science 3039:679-686.
Lobry, JR, N Sueoka. 2002. Asymmetric directional mutation pressures in bacteria. Genome Biol 3:RESEARCH0058.
Lucks, JB, DR Nelson, GR Kudla, JB Plotkin. 2008. Genome landscapes and bacteriophage codon usage. PLoS Comput Biol 4:e1000001.
Mandal, N, RK Ghosh. 1988. Characterization of the phage-specific transfer RNA molecules coded by cholera phage phi 149. Virology 166:583-585.
Marin, A, X Xia. 2008. GC skew in protein-coding genes between the leading and lagging strands in bacterial genomes: new substitution models incorporating strand bias. J Theor Biol 253:508-513.
Matsuzaki, S, M Rashel, J Uchiyama, et al. 2005. Bacteriophage therapy: a revitalized therapy against bacterial infectious diseases. J Infect Chemother 11:211-219.
McNair, K, BA Bailey, RA Edwards. 2012. PHACTS, a computational approach to classifying the lifestyle of phages. Bioinformatics 28:614-618.
Mrazek, J, S Karlin. 1998. Strand compositional asymmetry in bacterial and large viral genomes. Proc Natl Acad Sci U S A 95:3720-3725.
Muto, A, S Osawa. 1987. The guanine and cytocine content of genomic DNA and bacterial evolution. Proceedings of the National Academy of Sciences, USA 84:166-169.
NCBI. 2013. The Genetic Codes.

Ngumbela, KC, KP Ryan, R Sivamurthy, MA Brockman, RT Gandhi, N Bhardwaj, DG Kavanagh. 2008. Quantitative Effect of Suboptimal Codon Usage on Translational Efficiency of mRNA Encoding HIV-1 *gag* in Intact T Cells. PLoS One 3:e2356.
Ninio, J. 1975. Kinetic amplification of enzyme discrimination. Biochimie 57:587-595.

Ogle, JM, DE Brodersen, WM Clemons, Jr., MJ Tarry, AP Carter, V Ramakrishnan. 2001. Recognition of cognate transfer RNA by the 30S ribosomal subunit. Science 292:897-902.
Ogle, JM, FV Murphy, MJ Tarry, V Ramakrishnan. 2002. Selection of tRNA by the ribosome requires a transition from an open to a closed form. Cell 111:721-732.
Osawa, S, A Muto, TH Jukes, T Ohama. 1990. Evolutionary changes in the genetic code. Proc Biol Sci 241:19-28.
Palidwor, GA, TJ Perkins, X Xia. 2010. A general model of codon bias due to GC mutational bias. PLoS One 5:e13431.
Pavon-Eternod, M, A David, K Dittmar, P Berglund, T Pan, JR Bennink, JW Yewdell. 2013. Vaccinia and influenza A viruses select rather than adjust tRNAs to optimize translation. Nucleic Acids Res 41:1914-1921.
Percudani, R, A Pavesi, S Ottonello. 1997. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. J Mol Biol 268:322-330.
Plotkin, JB, H Robins, AJ Levine. 2004. Tissue-specific codon usage and the expression of human genes. Proc Natl Acad Sci U S A 101:12588-12591.
Ran, W, PG Higgs. 2012. Contributions of speed and accuracy to translational selection in bacteria. PLoS One 7:e51652.
Ranjan, A, AS Vidyarthi, R Poddar. 2007. Evaluation of codon bias perspectives in phage therapy of Mycobacterium tuberculosis by multivariate analysis. In Silico Biology 7:423-431.
Rice, P, I Longden, A Bleasby. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276-277.
Robinson, M, R Lilley, S Little, JS Emtage, G Yarranton, P Stephens, A Millican, M Eaton, G Humphreys. 1984. Codon usage can affect efficiency of translation of genes in Escherichia coli. Nucleic Acids Res 12:6663-6671.
Rohwer, F, R Edwards. 2002. The Phage Proteomic Tree: a genome-based taxonomy for phage. J Bacteriol 184:4529-4535.
Sahu, K, SK Gupta, S Sau, TC Ghosh. 2005. Comparative analysis of the base composition and codon usages in fourteen mycobacteriophage genomes. Journal of Biomolecular Structure & Dynamics 23:63-71.
SAS Institute. 1994. SAS/STAT User's Guide - Volume 2, GLM-VARCOMP.
Sau, K. 2007. Studies on synonymous codon and amino acid usages in Aeromonas hydrophila phage Aeh1: architecture of protein-coding genes and therapeutic implications. J Microbiol Immunol Infect 40:24-33.
Sau, K, A Deb. 2009. Temperature influences synonymous codon and amino acid usage biases in the phages infecting extremely thermophilic prokaryotes. In Silico Biology 9:1-9.
Sau, K, SK Gupta, S Sau, TC Ghosh. 2005. Synonymous codon usage bias in 16 Staphylococcus aureus phages: implication in phage therapy. Virus Res 113:123-131.
Sau, K, SK Gupta, S Sau, SC Mandal, TC Ghosh. 2007. Studies on synonymous codon and amino acid usage biases in the broad-host range bacteriophage KVP40. J Microbiol 45:58-63.
Schattner, P, AN Brooks, TM Lowe. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res 33:W686-689.
Shackelton, LA, EC Holmes. 2006. Phylogenetic evidence for the rapid evolution of human B19 erythrovirus. J Virol 80:3666-3669.

Shackelton, LA, CR Parrish, EC Holmes. 2006. Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. J Mol Evol 62:551-563.
Shackelton, LA, CR Parrish, U Truyen, EC Holmes. 2005. High rate of viral evolution associated with the emergence of carnivore parvovirus. Proceedings of the National Academy of Sciences of the United States of America 102:379-384.
Sharp, PM, E Cowe, DG Higgins, DC Shields, KH Wolfe, F Wright. 1988. Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity. Nucleic Acids Res 16:8207-8211.
Sharp, PM, KM Devine. 1989. Codon usage and gene expression level in Dictyostelium discoideum: highly expressed genes do 'prefer' optimal codons. Nucleic Acids Res 17:5029-5039.
Sharp, PM, WH Li. 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281-1295.
Sharp, PM, MS Rogers, DJ McConnell. 1984. Selection pressures on codon usage in the complete genome of bacteriophage T7. J Mol Evol 21:150-160.
Sharp, PM, TM Tuohy, KR Mosurski. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res 14:5125-5143.
Skiena, SS. 2001. Designing better phages. Bioinformatics 17 Suppl 1:S253-261.
Skurnik, M, M Pajunen, S Kiljunen. 2007. Biotechnological challenges of phage therapy. Biotechnol Lett 29:995-1003.
Steitz, TA. 2008. A structural understanding of the dynamic ribosome machine. Nat Rev Mol Cell Biol 9:242-253.
Sueoka, N. 1961. Correlation bewteen base composition of deoxyribonucleic acid and amino acid composition of proteins. Proceedings of the National Academy of Sciences, USA 47:1141-1149.
Sueoka, N. 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40:318-325.
Sun, XY, Q Yang, X X. 2013. An Improved Implementation of Effective Number of Codons (N_c). Molecular Biology and Evolution 30:191-196.
Timms, AR, J Cambray-Young, AE Scott, et al. 2010. Evidence for a lineage of virulent bacteriophages that target Campylobacter. BMC Genomics 11:214.
Tuller, T, YY Waldman, M Kupiec, E Ruppin. 2010. Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci U S A 107:3645-3650.
Umemura, T, Y Tanaka, K Kiyosawa, HJ Alter, JW Shih. 2002. Observation of positive selection within hypervariable regions of a newly identified DNA virus (SEN virus)(1). FEBS Lett 510:171-174.
Urbina, D, B Tang, PG Higgs. 2006. The response of amino acid frequencies to directional mutation pressure in mitochondrial genome sequences is related to the physical properties of the amino acids and to the structure of the genetic code. Journal of Molecular Evolution 62:340-361.
van Vliet, F, M Couturier, L Desmet, M Faelen, A Toussaint. 1978. Virulent mutants of temperate phage Mu-1. Molecular and General Genetics 160:195-202.
van Weringh, A, M Ragonnet-Cronin, E Pranckeviciene, M Pavon-Eternod, L Kleiman, X Xia. 2011. HIV-1 Modulates the tRNA Pool to Improve Translation Efficiency. Molecular Biology and Evolution 28:1827-1834.

Villegas, A, YM She, AM Kropinski, et al. 2009. The genome and proteome of a virulent Escherichia coli O157:H7 bacteriophage closely resembling Salmonella phage Felix O1. Virol J 6:41.
Woo, PC, BH Wong, Y Huang, SK Lau, KY Yuen. 2007. Cytosine deamination and selection of CpG suppressed clones are the two major independent biological forces that shape codon usage bias in coronaviruses. Virology 369:431-442.
Wright, F. 1990. The 'effective number of codons' used in a gene. Gene 87:23-29.
Xia, X. 1996. Maximizing transcription efficiency causes codon usage bias. Genetics 144:1309-1320.
Xia, X. 1998. How optimized is the translational machinery in Escherichia coli, Salmonella typhimurium and Saccharomyces cerevisiae? Genetics 149:37-44.
Xia, X. 2005. Mutation and selection on the anticodon of tRNA genes in vertebrate mitochondrial genomes. Gene 345:13-20.
Xia, X. 2007. An Improved Implementation of Codon Adaptation Index. Evolutionary Bioinformatics 3:53–58.
Xia, X. 2008. The cost of wobble translation in fungal mitochondrial genomes: integration of two traditional hypotheses. BMC Evol. Biol. 8:211.
Xia, X. 2012a. DNA replication and strand asymmetry in prokaryotic and mitochondrial genomes. Curr Genomics 13:16-27.
Xia, X. 2012b. Rapid evolution of animal mitochondria. In: RS Singh, J Xu, RJ Kulathinal, editors. Evolution in the fast lane: Rapidly evolving genes and genetic systems: Oxford University Press. p. 73-82
Xia, X. 2013a. Comparative genomics.: Springer.
Xia, X. 2013b. DAMBE5: A comprehensive software package for data analysis in molecular biology and evolution. Molecular Biology and Evolution 30:1720-1728.
Xia, X, H Huang, M Carullo, E Betran, EN Moriyama. 2007. Conflict between Translation Initiation and Elongation in Vertebrate Mitochondrial Genomes. PLoS One 2:e227.
Xia, X, G Palidwor. 2005. Genomic Adaptation to Acidic Environment: Evidence from *Helicobacter pylori*. Am. Nat. 166:776-784.
Xia, X, H Wang, Z Xie, M Carullo, H Huang, D Hickey. 2006. Cytosine usage modulates the correlation between CDS length and CG content in prokaryotic genomes. Molecular Biology and Evolution 23:1450-1454.
Xia, X, Z Xie. 2001. DAMBE: software package for data analysis in molecular biology and evolution. J Hered 92:371-373.
Xia, X, Q Yang. 2013. Cenancestor. In: S Maloy, K Hughes, editors. Encyclopedia of Genetics. San Diego: Academic Press. p. 493-494.
Xia, X, KY Yuen. 2005. Differential selection and mutation between dsDNA and ssDNA phages shape the evolution of their genomic AT percentage. BMC Genet 6:20.
Xu, Z, B Hao. 2009. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. Nucleic Acids Res 37:W174-178.

6 Appendix A - Supplemental Table

Table A. 1 – Genome details of dsDNA and ssDNA *E. coli* phages.

Phage Family	Phage Name	Phage accession	Strand Type	Genome Length	# CDS	GC%	TC Skew	# tRNA	Mean CAI	Median CAI	Std CAI	r (RSCU)
Myoviridae	Phage CC31	NC_014662	dsDNA	165540	279	39.92	0.1916	8	0.4996	0.4920	0.0805	0.2498
Myoviridae	Phage IME08	NC_014260	dsDNA	172253	253	39.59	0.2006	3	0.5011	0.4987	0.0809	0.3058
Myoviridae	Phage JS10	NC_012741	dsDNA	171451	265	39.52	0.2033	3	0.5065	0.5045	0.0810	0.2975
Myoviridae	Phage JS98	NC_010105	dsDNA	170523	266	39.51	0.2016	3	0.5094	0.5052	0.0808	0.2977
Myoviridae	Phage JSE	NC_012740	dsDNA	166418	277	40.51	0.1779	0	0.4979	0.4869	0.0875	0.4789
Myoviridae	Phage Mu	NC_000929	dsDNA	36717	55	52.05	-0.0251	0	0.5054	0.5041	0.0677	0.9207
Myoviridae	Phage P1	NC_005856	dsDNA	94800	110	47.31	0.0585	3	0.4533	0.4485	0.0666	0.8224
Myoviridae	Phage P2	NC_001895	dsDNA	33593	43	50.17	0.0220	0	0.4967	0.5128	0.0830	0.9011
Myoviridae	Phage P4	NC_001609	dsDNA	11624	14	49.53	-0.0009	0	0.4723	0.4918	0.0554	0.8287
Myoviridae	Phage Phi1	NC_009821	dsDNA	164270	276	40.50	0.1773	0	0.5012	0.4897	0.0870	0.4971
Myoviridae	Phage phiEcoM-GJ1	NC_010106	dsDNA	52975	75	44.02	0.1319	0	0.4915	0.4754	0.0842	0.3437
Myoviridae	Phage phiP27	NC_003356	dsDNA	42575	58	49.35	0.0613	2	0.4711	0.4695	0.0652	0.8296
Myoviridae	Phage RB16	NC_014467	dsDNA	176788	270	43.52	0.1311	2	0.5437	0.5417	0.0869	0.6884
Myoviridae	Phage RB49	NC_005066	dsDNA	164018	279	40.44	0.1794	0	0.5001	0.4870	0.0846	0.4917
Myoviridae	Phage RB69	NC_004928	dsDNA	167560	273	37.66	0.2409	2	0.4831	0.4741	0.0730	0.1957
Myoviridae	Phage SfV	NC_003444	dsDNA	37074	53	50.77	0.0174	0	0.4833	0.4943	0.0620	0.8750
Myoviridae	Phage T4	NC_000866	dsDNA	168903	278	35.30	0.2725	8	0.4532	0.4488	0.0713	0.1311
Myoviridae	Phage vB_EcoM-VR7	NC_014792	dsDNA	169285	293	40.33	0.1805	1	0.5032	0.5025	0.0853	0.4432
Myoviridae	Phage WV8	NC_012749	dsDNA	88486	140	38.89	0.2288	20	0.4692	0.4672	0.0717	0.1517
Myoviridae	Phage D108	NC_013594	dsDNA	37235	57	51.76	-0.0204	0	0.5051	0.5023	0.0724	0.9207
Myoviridae	Phage rv5	NC_011041	dsDNA	137947	233	43.56	0.1289	5	0.4698	0.4714	0.0738	0.3680
Myoviridae	Phage vB_EcoM_CBA120	NC_016570	dsDNA	157304	204	44.50	0.1121	3	0.4872	0.4730	0.0782	0.6297
Podoviridae	Phage 13a	NC_011045	dsDNA	38841	55	48.39	0.0434	0	0.4686	0.4672	0.0879	0.3181
Podoviridae	Phage 285P	NC_015249	dsDNA	39270	47	48.73	0.0395	0	0.5168	0.5171	0.0794	0.4573

Continued on next page

Continued from previous page

PhageFam	PhageName	Phage accession	Strand	Genome Length	Num CDS	GC%	TC Skew	NtRNA	Mean CAI	Median CAI	Std CAI	r (RSCU)
Podoviridae	Phage 933W	NC_000924	dsDNA	61670	80	49.37	0.0411	3	0.4546	0.4501	0.0634	0.8154
Podoviridae	Phage EcoDS1	NC_011042	dsDNA	39252	53	49.94	0.0047	0	0.4698	0.4663	0.0842	0.4021
Podoviridae	Phage K1-5	NC_008152	dsDNA	44385	52	45.25	0.1329	0	0.4793	0.4674	0.0818	0.2629
Podoviridae	Phage K1E	NC_007637	dsDNA	45251	62	45.05	0.1368	0	0.4684	0.4663	0.0818	0.2553
Podoviridae	Phage K1F	NC_007456	dsDNA	39704	43	49.78	0.0171	0	0.4859	0.4751	0.0794	0.2553
Podoviridae	Phage Min27	NC_010237	dsDNA	63395	83	49.50	0.0402	3	0.4549	0.4582	0.0642	0.8388
Podoviridae	Phage N4	NC_008720	dsDNA	70153	72	41.30	0.1679	0	0.4978	0.4933	0.0711	0.2661
Podoviridae	Phage Phieco32	NC_010324	dsDNA	77554	128	42.27	0.1453	1	0.4675	0.4565	0.0798	0.1706
Podoviridae	Phage T3	NC_003298	dsDNA	38208	55	49.90	0.0063	0	0.4981	0.5054	0.0832	0.5306
Podoviridae	Phage T7	NC_001604	dsDNA	39937	60	48.40	0.0387	0	0.4767	0.4677	0.0914	0.3274
Podoviridae	Phage VT2-Sakai	NC_000902	dsDNA	60942	83	49.91	0.0304	3	0.4508	0.4488	0.0614	0.8248
Podoviridae	Phage phiV10	NC_007804	dsDNA	39104	55	48.97	0.0382	0	0.4870	0.4929	0.0595	0.7308
Podoviridae	Phage BA14	NC_011040	dsDNA	39816	52	48.78	0.0413	0	0.5123	0.5240	0.1009	0.4504
Podoviridae	Phage Stx2-I	NC_003525	dsDNA	61765	166	49.39	0.0409	0	0.4218	0.4222	0.0685	0.8034
Siphoviridae	Phage BP-4795	NC_004813	dsDNA	57930	85	50.61	0.0268	0	0.4505	0.4487	0.0745	0.8049
Siphoviridae	Phage cdtI	NC_009514	dsDNA	47021	60	49.12	0.0618	0	0.4623	0.4616	0.0761	0.8307
Siphoviridae	Phage EPS7	NC_010583	dsDNA	111382	170	39.90	0.2011	0	0.4751	0.4737	0.0935	0.2583
Siphoviridae	Phage HK022	NC_002166	dsDNA	40751	57	49.48	0.0112	0	0.4473	0.4542	0.0583	0.7416
Siphoviridae	Phage HK97	NC_002167	dsDNA	39732	61	49.79	-0.0019	0	0.4614	0.4601	0.0634	0.7303
Siphoviridae	Phage JK06	NC_007291	dsDNA	46072	82	44.00	0.1047	0	0.4365	0.4283	0.0721	0.2565
Siphoviridae	Phage lambda	NC_001416	dsDNA	48502	73	49.86	0.0267	0	0.4579	0.4634	0.0781	0.8520
Siphoviridae	Phage N15	NC_001901	dsDNA	46375	60	51.17	-0.0173	0	0.4773	0.4668	0.0650	0.8955
Siphoviridae	Phage RTP	NC_007603	dsDNA	46219	75	44.28	0.1201	0	0.4686	0.4618	0.0868	0.2412
Siphoviridae	Phage SSL-2009a	NC_012223	dsDNA	39792	52	54.72	-0.0812	0	0.5183	0.5111	0.0816	0.7853
Siphoviridae	Phage T1	NC_005833	dsDNA	48836	78	45.55	0.0983	0	0.4684	0.4687	0.0784	0.4637
Siphoviridae	Phage T5	NC_005859	dsDNA	121750	162	39.27	0.2178	18	0.4436	0.4316	0.0927	0.2071
Siphoviridae	Phage TLS	NC_009540	dsDNA	49902	87	42.68	0.1657	0	0.4862	0.4909	0.0589	0.4734
Siphoviridae	Phage HK639	NC_016158	dsDNA	49576	76	52.45	-0.0506	0	0.4916	0.4832	0.0746	0.8225

Continued on next page

Continued from previous page

PhageFam	PhageName	Phage accession	Strand	Genome Length	Num CDS	GC%	TC Skew	N tRNA	Mean CAI	Median CAI	Std CAI	r (RSCU)
Siphoviridae	Phage HK75	NC_016160	dsDNA	36661	58	50.19	-0.0073	0	0.4620	0.4618	0.0673	0.7605
Siphoviridae	Stx1 converting bacterioPhage	NC_004913	dsDNA	59866	84	49.69	0.0370	0	0.4519	0.4488	0.0654	0.8108
Siphoviridae	Stx2-Phage II	NC_004914	dsDNA	62706	89	49.90	0.0293	3	0.4578	0.4614	0.0600	0.8143
Siphoviridae	Stx2-Phage 1717	NC_011357	dsDNA	62148	77	50.92	0.0084	0	0.4544	0.4569	0.0634	0.8335
Siphoviridae	Stx2-Phage 86	NC_008464	dsDNA	60238	81	49.07	0.0493	3	0.4499	0.4629	0.0591	0.7787
Siphoviridae	Phage SPC35	NC_015269	dsDNA	118351	145	39.39	0.2110	18	0.4586	0.4465	0.0926	0.2205
Siphoviridae	Phage bV_EcoS_AKFV33	NC_017969	dsDNA	108853	160	38.95	0.2198	19	0.4567	0.4505	0.0904	0.2225
Tectiviridae	Phage PRD1	NC_001421	dsDNA	14927	31	48.10	0.0643	0	0.4714	0.4693	0.0565	0.6084
Inoviridae	Phage I2-2	NC_001332	ssDNA	6744	9	42.72	0.2321		0.5384	0.5133	0.0990	0.3572
Inoviridae	Phage If1	NC_001954	ssDNA	8454	10	43.71	0.1645		0.4255	0.3993	0.0617	0.3922
Inoviridae	Phage Ike	NC_002014	ssDNA	6883	10	40.55	0.2893		0.5190	0.5182	0.0609	0.3600
Inoviridae	Phage M13	NC_003287	ssDNA	6407	10	40.74	0.2630		0.5006	0.4704	0.0709	0.2700
Microviridae	Phage alpha3	NC_001330	ssDNA	6087	10	45.18	0.1798		0.4839	0.4924	0.0860	0.3709
Microviridae	Phage G4	NC_001420	ssDNA	5577	11	45.69	0.0213		0.4591	0.4417	0.0591	0.3036
Microviridae	Phage ID18	NC_007856	ssDNA	5486	11	45.22	0.0593		0.4629	0.4497	0.0605	0.2855
Microviridae	Phage ID2 Moscow/ID/2001	NC_007817	ssDNA	5486	11	45.75	0.0014		0.4551	0.4585	0.0638	0.2757
Microviridae	Phage phiX174	NC_001422	ssDNA	5386	11	44.76	0.1855		0.4720	0.4685	0.0873	0.2775
Microviridae	Phage St-1	NC_012868	ssDNA	6094	11	45.21	0.1829		0.4714	0.4870	0.0814	0.3555
Microviridae	Phage WA13	NC_007821	ssDNA	6068	10	44.83	0.1993		0.4941	0.5184	0.0928	0.3521

- refers to number, Std- standard deviation