The Effect of Mutation and Selection on Codon Adaptation in *Escherichia coli* Bacteriophage

Shivapriya Chithambaram,* Ramanandan Prabhakaran,* and Xuhua Xia*,†,1
*Department of Biology and Center for Advanced Research in Environmental Genomics, University of Ottawa, Ottawa, Ontario, K1N 6N5, Canada, and †Ottawa Institute of Systems Biology, Ottawa, Ontario, K1H 8M5 Canada

ABSTRACT Studying phage codon adaptation is important not only for understanding the process of translation elongation, but also for reengineering phages for medical and industrial purposes. To evaluate the effect of mutation and selection on phage codon usage, we developed an index 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 occur more frequently in single-stranded DNA (ssDNA) phages than in double-stranded DNA (dsDNA) phages and affect not only synonymous codon usage, but also nonsynonymous substitutions at second codon positions, especially in ssDNA phages. The host translation machinery affects codon adaptation in both dsDNA and ssDNA phages, with a stronger effect on dsDNA phages than on ssDNA phages. Strand asymmetry with the associated local variation in mutation bias can significantly interfere with codon adaptation in both dsDNA and ssDNA phages.

CODON adaptation has been well documented in bacterial and fungal genomes (Gouy and Gautier 1982; Ikemura 1981, 1992; Xia 1998) as well as in mitochondrial genomes in vertebrates (Xia 2005; Xia *et al.* 2007) and fungi (Carullo and 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 *et al.* 1996; Ngumbela *et al.* 2008) or transgenic genes (Hernan *et al.* 1992; Kleber-Janke and Becker 2000; Koresawa *et al.* 2000). Studies on codon–anticodon adaptation have progressed in theoretical elaboration (Bulmer 1987, 1991; Xia 1998, 2008; Higgs and Ran 2008; Jia and Higgs 2008; Palidwor *et al.* 2010) as well as in critical tests of alternative theoretical predictions (Xia 1996, 2005; Carullo and Xia 2008; van Weringh *et al.* 2011).

Codon–anticodon adaptation has been documented in bacteriophage (referred to as phage hereafter), partly because several phage species have been used to treat human infections (Sau *et al.* 2005; Ranjan *et al.* 2007; Sau 2007; Abedon *et al.* 2011) or remove infectious biofilms (Azeredo and Sutherland 2008) and need to be reengineered to improve translation efficiency. While phage codon adaptation is shaped mainly by mutation and transfer (t)RNA-mediated selection, previous studies (Grosjean *et al.* 1978; Gouy 1987; Kunisawa *et al.* 1998; Sahu *et al.* 2005; Carbone 2008; Lucks *et al.* 2008) have focused mainly on the tRNA-mediated selection on codon usage and have not assessed quantitatively the joint 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 and Miller 1980; Lindahl 1993; Xia and Yuen 2005). In particular, the spontaneous deamination rate is ~100 times higher in ssDNA than in double-stranded DNA (dsDNA) based on experimental evidence (Frederico *et al.* 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 and Yuen 2005; Shackelton and Holmes 2006; Duffy and Holmes 2008, 2009). Oxidative deamination leading to high

doi: 10.1534/genetics.114.162842
Manuscript received January 7, 2014; accepted for publication February 23, 2014; published Early Online February 28, 2014.
Available freely online through the author-supported open access option.
[1] Corresponding author: Department of Biology and Center for Advanced Research in Environmental Genomics, University of Ottawa, 30 Marie Curie, P.O. Box 450, Station A, Ottawa, ON, K1N 6N5, Canada. E-mail: xxia@uottawa.ca

C→U/T transitional mutation rates has been reported in ssDNA phage M13 (Kreutzer and Essigmann 1998).

The Effect of C→T Mutation Bias

When selection is absent, if the C→T mutation bias experienced by a phage genome is strong enough to overcome 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 tend to have the same proportion of U-ending codons; *i.e.*,

$$P_{\mathrm{U}.i} = \frac{N_{\mathrm{U}.i}}{N_{\mathrm{U}.i} + N_{\mathrm{C}.i}} = B_{\mathrm{C}\to\mathrm{T}}, \tag{1}$$

where $N_{\mathrm{U}.i}$ and $N_{\mathrm{C}.i}$ are the numbers of codons ending with U or C, respectively, in codon family i, and $B_{\mathrm{C}\to\mathrm{T}}$ is a constant representing C→T mutation bias (being 0.5 when there is no C→T mutation bias). When $B_{\mathrm{C}\to\mathrm{T}}$ increases, P_{U} for all codon families will tend to increase synchronously if not checked by selection. Equation 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, $B_{\mathrm{C}\to\mathrm{T}}$ can be approximated simply as the average of $P_{\mathrm{U}.i}$ values for all Y-ending codon families in the viral protein-coding genes; *i.e.*,

$$B_{\mathrm{C}\to\mathrm{T}} = \overline{P}_{\mathrm{U}} = \frac{\sum_{i=1}^{N_{\mathrm{Y}}} P_{\mathrm{U}.i}}{N_{\mathrm{Y}}}, \tag{2}$$

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

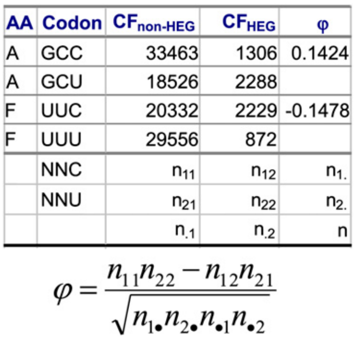
Some ssDNA phages have high $\overline{P}_{\mathrm{U}}$ values; *e.g.*, $\overline{P}_{\mathrm{U}}$ at the third codon position of *Chlamydia* phage Chp1 (Microviridae, NC_001741) is 0.9518, with U-ending codons being invariably the most frequent in all Y-ending or N-ending codon families. Some dsDNA phages can also have high $\overline{P}_{\mathrm{U}}$, *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.

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 refer to 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 do we measure U friendliness (*i.e.*, selection in favor of U-ending codons)? We develop a simple index, numerically illustrated in Figure 1, based on the comparison of codon frequency (CF) between highly expressed host genes (HEGs) and all other genes (non-HEGs), designated by $\mathrm{CF}_{\mathrm{HEG}}$ and $\mathrm{CF}_{\mathrm{non\text{-}HEG}}$, 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, with only HEGs from *Escherichia coli*, 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. The mechanistic explanation for this is that GCC can be decoded only by $\mathrm{tRNA}^{\mathrm{Ala/GGC}}$, whereas GCU can be decoded by both $\mathrm{tRNA}^{\mathrm{Ala/GGC}}$ and $\mathrm{tRNA}^{\mathrm{Ala/UGC}}$, where the wobble U in the anticodon is modified to $\mathrm{cmo}^5\mathrm{U}$, which pairs efficiently with U at the third codon position (Mitra *et al.* 1977, 1979; Nasvall *et al.* 2007). 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 *E. coli* HEGs are included. This is also expected because these codons are decoded by $\mathrm{tRNA}^{\mathrm{Phe/GAA}}$, which prefers UUC over UUU. The observation that UUC is preferred by HEGs suggests that the Phe codon family is not U-friendly (it is C 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 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 $P_{\mathrm{U}.i}$.

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 from HEGs are 80 and 90, respectively, but those for all CDSs are 200 and 600, respectively. A proper interpretation of this scenario is that extremely high T-biased mutation leads to the dominance of NNU codons. However, the host translation machinery prefers C-ending codons and this selection acts against the T-biased mutation so that codon usage in HEGs is not as U-biased as that in all CDSs. If we have codon usage of only HEGs, we may conclude that the host translation machinery prefers U-ending codons.



AA	Codon	$CF_{non\text{-}HEG}$	CF_{HEG}	φ
A	GCC	33463	1306	0.1424
A	GCU	18526	2288	
F	UUC	20332	2229	-0.1478
F	UUU	29556	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 1 Rationale of using the ϕ-coefficient as a proxy for U friendliness, based on the codon frequencies (CF) between highly expressed genes (HEGs) and non-HEGs from *E. coli*. ϕ can take values within the range between −1 and 1.

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 Equation 1 to include the effect of selection on $P_{U.i}$; *i.e.*,

$$P_{U.i} = B_{C\rightarrow T} + b\varphi_i. \quad (3)$$

Because $P_{U.i}$ can be readily computed from viral protein-coding genes, and ϕ_i can be derived from codon frequencies of host HEGs and non-HEGs (Figure 1), we can use Equation 3 to quantify the relative importance of mutation ($B_{C\rightarrow T}$) 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 $B_{C\rightarrow T}$, and we will observe little variation in $P_{U.i}$ values among different codon families. In contrast, with increasing intensity of selection (a large b) or increasing variation in preference of U-ending codons by *E. coli* (*i.e.*, large variation in ϕ), $P_{U.i}$ will become more dependent on $b\phi_i$. Similarly, if $B_{C\rightarrow T}$ becomes very large (*i.e.*, very strong mutation bias), then $b\phi_i$ 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 >1 or <0, so it will asymptotically approach 1 with increasing ϕ_i and approach 0 with decreasing ϕ_i. This implies a sigmoidal relationship between P_U and ϕ. For this reason, we have also fitted the following sigmoid function,

$$P_{U.i} = \frac{1}{1 + Ce^{-D\varphi_i}}, \quad (4)$$

where parameters C and D are constants. The maximum and minimum values for P_U, according to Equation 4, are 1 and 0, respectively. When $\phi = 0$ or $D = 0$, the expected P_U is $1/(1 + C)$, which is equivalent to $B_{C\rightarrow T}$ in Equation 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. D measures the benefit of codon adaptation for the phage. If $D = 0$ (*i.e.*, codon adaptation is not important for the phage), then which codon is favored by the host machinery (measured by ϕ) is irrelevant to phage codon usage. If D is very large, then even a codon that is weakly favored by the host will be strongly favored by the phage.

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 in phages. 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 exhibits good codon adaptation to host species B, although it is preadaptation that is applicable here.

We use Equation 3 to characterize mutation bias and selection intensity based on 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. Some *E. coli* phages such as phage PRD1 whose close relatives are all parasitizing gram-positive bacteria may have recently invaded *E. coli* and have codon usage highly different from that of *E. coli* HEGs. Strand asymmetry with the associated local variation in mutation bias (U-biased in one-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.

Materials and Methods

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 the "FEATURES" table, including 60 dsDNA phages and 11 ssDNA phages. The CDSs and codon usage data, as well as three codon positions, were extracted using DAMBE (Xia 2013b). All phage genomes were searched for encoded tRNAs by using the tRNAscan-SE Search Server (Schattner *et al.* 2005). The local TC skew plot, with the TC skew computed as $(N_T - N_C)/(N_T + N_C)$, where N_i is the number of nucleotides i along a moving window, was 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 the 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 *et al.* 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 codon usage in highly expressed genes among different *E. coli* genomes.

Indexes of codon usage bias

While we mainly focus on modeling mutation and selection on P_U in Equations 3 and 4, two indexes of codon usage bias were used to aid in the interpretation of the results: the codon adaptation index (CAI) (Sharp and Li 1987) with the improved implementation (Xia 2007) and the effective number of codons (N_c) (Wright 1990) with the improved implementation (Sun *et al.* 2013). All these indexes were computed using DAMBE (Xia 2013b). For computing the phage CAI, the host highly expressed genes are used as the reference set of genes. Only CDSs with at least 33 codons (99 nt) are included in computing the indexes of codon usage bias to alleviate stochastic noise in computing these indexes with few codons.

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 and Hao 2009) that does not require aligned sequences but implicitly assumes the sharing of ancestral peptides as phylogenetic signals. The method uses amino acid sequences and is conceptually based on the sharing of ancient peptides that give individual evolutionary lineages their uniqueness. Computationally, the method is built upon the similarities in the sharing of words of length k ($a_1a_2 \dots a_k$) after subtracting its random expectation based on the frequencies of $a_1a_2 \dots a_{k-1}$, $a_2a_3 \dots a_k$, and $a_2a_3 \dots a_{k-1}$. The CVTree method has been implemented in the most recent version of DAMBE (Xia 2013b). We used a k value of 5, which has been recommended for viral genomes (Xu and 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.

Table 1 Codon frequencies (CF) for Y-ending codons in *E. coli*, compiled for highly expressed genes (HEG) and all other genes (non-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 corresponds to greater preference of U-ending codons over C-ending codons)

AA	Codon	$CF_{non\text{-}HEG}$	CF_{HEG}	tRNA	ϕ
A	GCC	33,463	1,306	2	0.1424
A	GCT	18,526	2,288		
C	TGC	8,397	475	1	−0.0362
C	TGT	6,802	270		
D	GAC	23,226	2,786	3	−0.0993
D	GAT	41,472	2,345		
F	TTC	20,332	2,229	2	−0.1478
F	TTT	29,556	872		
G	GGC	37,418	2,987	4	0.0566
G	GGT	30,154	3,583		
H	CAC	12,144	1,160	1	−0.1331
H	CAT	17,170	477		
I	ATC	30,787	3,488	3	−0.1232
I	ATT	39,788	1,640		
L	CTC	14,591	541	1	−0.0353
L	CTT	14,679	357		
N	AAC	26,674	2,832	4	−0.1512
N	AAT	23,652	539		
P	CCC	7,443	38	1	0.1032
P	CCT	9,235	343		
R	CGC	28,473	1,530		0.1011
R	CGT	25,528	2,995	3[a]	
S	AGC	20,868	1,015	1	−0.0842
S	AGT	11,802	168		
S	TCC	10,649	1,110	2	0.0327
S	TCT	10,217	1,320		
T	ACC	29,335	2,533	2	0.0408
T	ACT	10,950	1,286		
V	GTC	19,972	824	2	0.1262
V	GTT	22,297	2,669		
Y	TAC	15,094	1,569	3	−0.1122
Y	TAT	21,207	865		

P-values from a chi-square test of 2 × 2 contingency tables, with the null hypothesis that $\phi = 0$, are all <0.0001.

[a] The anticodon has a wobble A modified to inosine.

Results and Discussion

Codon preference by the E. coli translation machinery: ϕ

The ϕ-values for *E. coli* Y-ending codons are generally small, ranging from −0.1512 to 0.1424 (Table 1). Among the 16 Y-ending codon families and subfamilies, 7 are U-friendly (with $\phi > 0$, with the mean = 0.0861) and 9 are U-hostile (with $\phi < 0$, with the mean = −0.1025). 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 ($P_{U.E.coli} = 0.4421$). Increased C→T mutation bias will improve codon adaptation for the

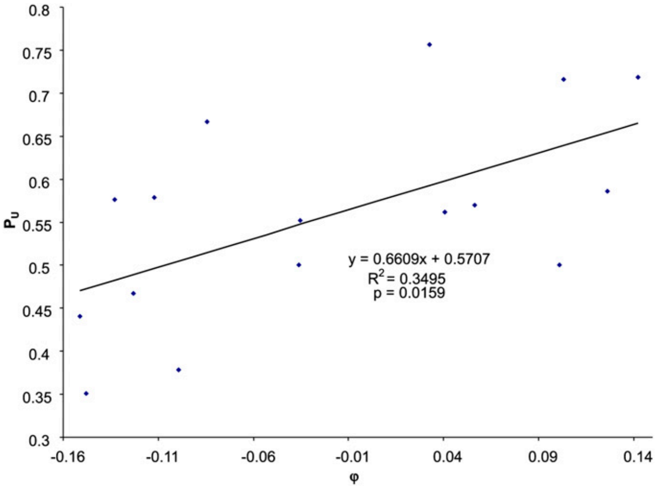


Figure 2 Relationship between P_U (the proportion of U-ending codons in Y-ending codon families) and ϕ (selection in favor of U-ending codons), based on codon usage data from *E. coli* Enterobacteria phage G4 (NC_001420). Also shown is the linear fit to the data. Applying the sigmoid function in Equation 4 generated effectively the same predicted values.

7 U-friendly codon families, but will lead to deterioration in the 9 U-hostile codon families (Table 1).

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 2). Fitting the linear model in Equation 3 results in $B_{C\rightarrow T} = 0.5707$, and $b = 0.6609$, with the relationship being statistically significant ($P = 0.0159$). Fitting the sigmoid function in Equation 4 yields $C = 0.7475$, $D = 2.7241$, and $1/(1 + C) = 0.5722$, which is equivalent to $B_{C\rightarrow T}$ in the linear model, *i.e.*, both being the expected P_U value when $D\phi$ in Equation 4 is zero (*i.e.*, no selection on phage codon usage in Y-ending codon families). The predicted values from the linear model in Equation 3 and the nonlinear model in Equation 4 are identical to the first two digits after the decimal point, indicating sufficiency of the linear model.

The estimated $B_{C\rightarrow T}$ from applying the regression model in Equation 3 to all 11 ssDNA Enterobacteria phages parasitizing *E. coli* varies from 0.5443 to 0.7419 (Table 2). These would be the P_U values when selection mediated by the host translation machinery is absent. With the slopes in Table 2, the effect of selection on viral P_U values is small relative to $B_{C\rightarrow T}$. We thus expect the estimated $B_{C\rightarrow T}$ values to be close to the empirical $\overline{P}_U$ values defined in Equation 2, which is true (Figure 3).

The standard error associated with the $B_{C\rightarrow T}$ values is on the order of 0.02 (not shown), so that $B_{C\rightarrow T}$ values in Table 2 are all significantly greater than the observed $\overline{P}_U$ (= 0.4421) in *E. coli* HEGs. If we assume that the codon usage of *E. coli* HEGs represents the optimum achievable given the counterbalance between mutation and selection, then the large $B_{C\rightarrow T}$ values in Table 2 suggest that C→T-biased mutation in ssDNA has shifted the codon usage of ssDNA phages away from the optimum.

$B_{C\rightarrow T}$ has a strong effect on the effective number of codons (N_c), as expected. N_c is at its maximum when $B_{C\rightarrow T} \sim 0.5$, but decreases sharply as $B_{C\rightarrow T}$ 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 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

Table 2 Results of fitting the linear regression model in Equation 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 ($B_{C\rightarrow T}$) 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	$B_{C\rightarrow T}$	b	R	P
Phage alpha3	NC_001330	6087	10	0.7022	0.3790	0.3936	0.1315
Phage G4	NC_001420	5577	11	0.5707	0.6609	0.5912	0.0159
Phage ID18	NC_007856	5486	11	0.5876	0.6551	0.5105	0.0433
Phage ID2	NC_007817	5486	11	0.5443	0.5820	0.5055	0.0458
Phage phiX174	NC_001422	5386	11	0.6965	0.2840	0.2969	0.2641
Phage St-1	NC_012868	6094	11	0.6881	0.2895	0.3048	0.2511
Phage WA13	NC_007821	6068	10	0.7122	0.2190	0.2354	0.3801
Phage I2-2	NC_001332	6744	9	0.6987	0.3301	0.2649	0.3214
Phage If1	NC_001954	8454	10	0.6551	−0.0763	−0.0819	0.7629
Phage Ike	NC_002014	6883	10	0.7419	0.1614	0.1097	0.6858
Phage M13	NC_003287	6407	10	0.7390	0.1717	0.1899	0.4812

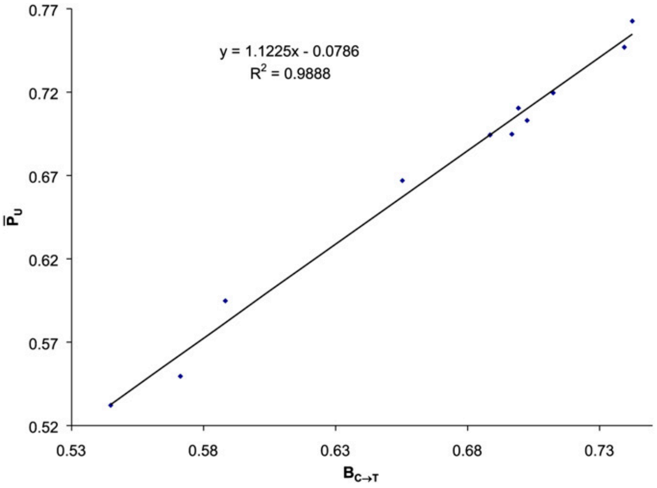


Figure 3 The average $\overline{P}_U$ defined in Equation 2 is similar to $B_{C\to T}$ estimated from fitting the linear model in Equation 3, based on 11 ssDNA Enterobacteria phages parasitizing *E. coli* (Table 2).

empirical data (Figure 4), although there is one outlying point (Enterobacteria phage If1; NC_001954) for which we 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 diamonds in Figure 4) have codon usages significantly correlated with each other, which suggests that they might be phylogenetically related. The tree built with the CVTree algorithm (Xu and Hao 2009) implemented in DAMBE (Xia 2013b) does cluster these four species, all belonging to Inoviridae, into a monophyletic taxon (Figure 5). Other viral proteomic trees (Rohwer and Edwards 2002; Edwards and Rohwer 2005) also group these four *E. coli* ssDNA phages into the same clade. The other seven ssDNA phages, all belonging to Microviridae, are also clustered into a monophyletic taxon (Figure 5).

Because of the phylogenetic structure (Figure 5), one may argue that the 11 points are not statistically independent and question the validity of using conventional regression to test the significance of the negative association between R and $B_{C\to T}$. For example, the ancestor of the seven phages in Microviridae (colored red in Figure 5) may have codon usage similar to that in *E. coli*, which was then inherited by all seven descendant phage lineages. Similarly, the ancestor of the four phage species in Inoviridae (colored blue in Figure 5) may have codon usage different from that of *E. coli*, which was then inherited by its four descendant phage lineages. Thus, we would, in an extreme case, have only two data points. To overcome this problem, we used the subtree for the 11 species in Figure 5 and performed independent-contrasts analysis (Felsenstein 1985, 2004, pp. 435–443) implemented in DAMBE (Xia 2013a, pp. 24–29; Xia 2013b) and found the negative association still significant ($P < 0.05$).

We may conclude from the results above that the overall effect of selection in ssDNA phage is statistically significant as the mean of the 11 b values is significantly >0 (mean b = 0.3324, SE = 0.0685, t = 4.8546, degrees of freedom (d.f.) = 10, P = 0.0007). However, when the false discovery rate method (Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001) is used to control for type I error rate involving multiple comparisons, none of the 11 individual P-values is statistically significant at the 0.05 significance level. The false discovery rate method has been numerically illustrated (Xia 2012b) and implemented in DAMBE (Xia 2013b).

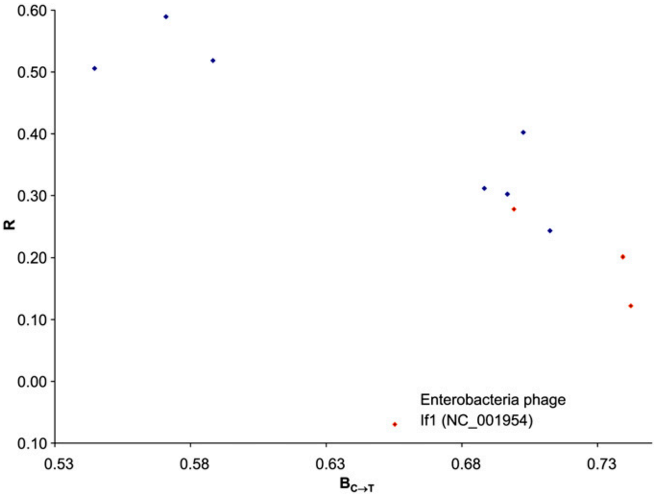


Figure 4 The correlation (R) between P_U and ϕ decreases with increasing $B_{C\to T}$. The outline point is *E. coli* Enterobacteria phage If1 (NC_001954). The negative association is statistically significant (P = 0.0292 with the outlying point included). The four red circles form a monophyletic taxon and the rest form another monophyletic taxon (Figure 5).

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 6). The estimated $B_{C\to T}$ spans a wide range (Table 3), but on average is significantly smaller than that of ssDNA phages (t-test, t = 2.1379, d.f. = 69, two-tailed P = 0.0361), suggesting a weaker effect of C→T mutation on codon usage in dsDNA phages than in ssDNA phages. The b values also vary substantially (Table 3). For example, phage BP-4795 (NC_004813), phage cdtI (NC_009514), and 10 other dsDNA phages have negative slopes between P_U and ϕ (Table 3), contrary to what we would have predicted based on codon adaptation. While it is easy to understand why phages should exhibit codon adaptation (with a positive slope b) to the host, it is puzzling why some dsDNA phages do not evolve codon usage similar to that of the host.

We noted that all phages that have negative correlation between P_U and ϕ share similar codon usages. 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

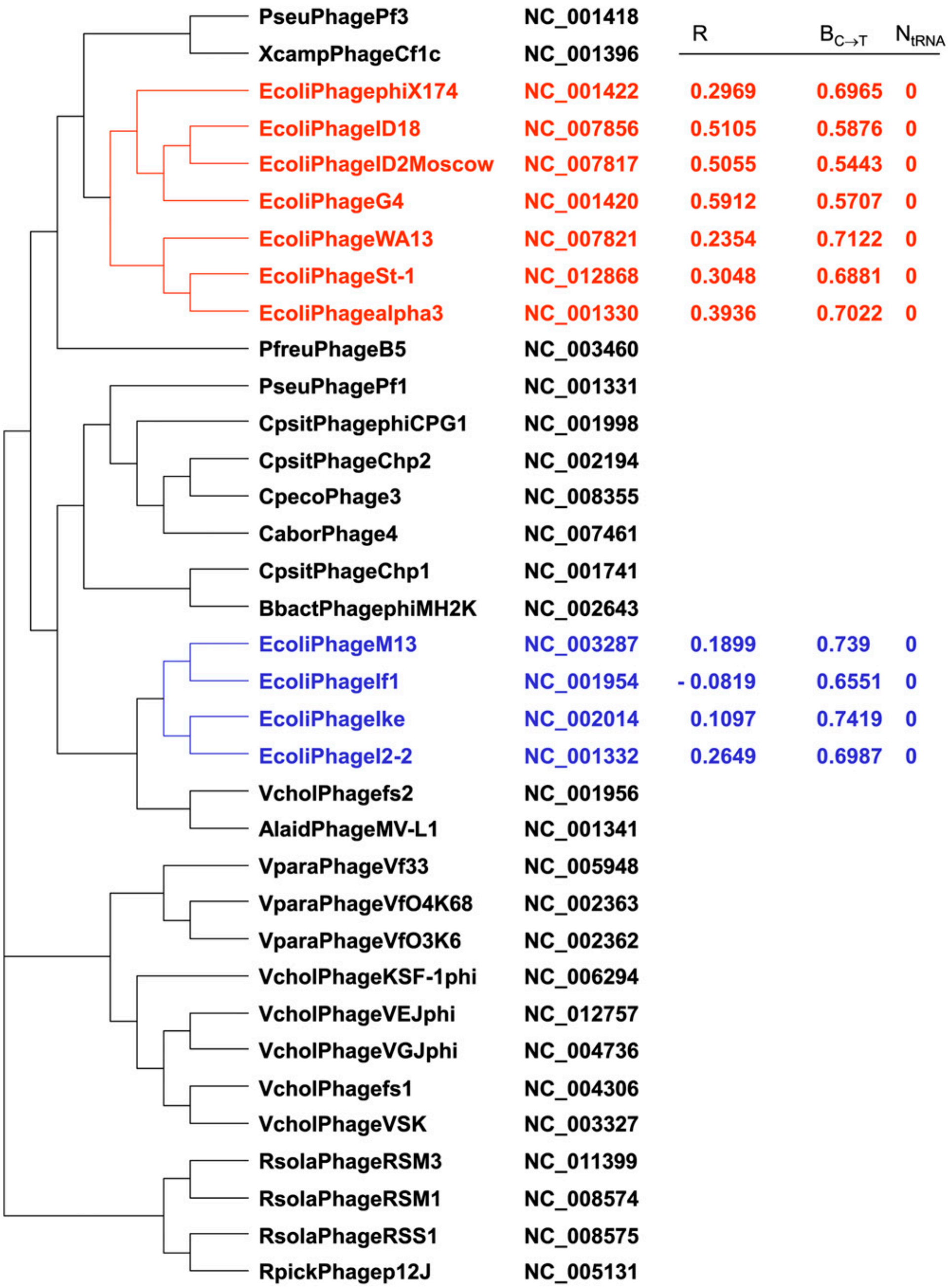

R
$B_{C\to T}$
N_{tRNA}
PseuPhagePf3 NC_001418
XcampPhageCf1c NC_001396
EcoliPhagephiX174 NC_001422 0.2969 0.6965 0
EcoliPhageID18 NC_007856 0.5105 0.5876 0
EcoliPhageID2Moscow NC_007817 0.5055 0.5443 0
EcoliPhageG4 NC_001420 0.5912 0.5707 0
EcoliPhageWA13 NC_007821 0.2354 0.7122 0
EcoliPhageSt-1 NC_012868 0.3048 0.6881 0
EcoliPhagealpha3 NC_001330 0.3936 0.7022 0
PfreuPhageB5 NC_003460
PseuPhagePf1 NC_001331
CpsitPhagephiCPG1 NC_001998
CpsitPhageChp2 NC_002194
CpecoPhage3 NC_008355
CaborPhage4 NC_007461
CpsitPhageChp1 NC_001741
BbactPhagephiMH2K NC_002643
EcoliPhageM13 NC_003287 0.1899 0.739 0
EcoliPhageIf1 NC_001954 - 0.0819 0.6551 0
EcoliPhageIke NC_002014 0.1097 0.7419 0
EcoliPhageI2-2 NC_001332 0.2649 0.6987 0
VcholPhagefs2 NC_001956
AlaidPhageMV-L1 NC_001341
VparaPhageVf33 NC_005948
VparaPhageVfO4K68 NC_002363
VparaPhageVfO3K6 NC_002362
VcholPhageKSF-1phi NC_006294
VcholPhageVEJphi NC_012757
VcholPhageVGJphi NC_004736
VcholPhagefs1 NC_004306
VcholPhageVSK NC_003327
RsolaPhageRSM3 NC_011399
RsolaPhageRSM1 NC_008574
RsolaPhageRSS1 NC_008575
RpickPhagep12J NC_005131

Figure 5 Phylogenetic tree of ssDNA phages reconstructed by using the CVTree method (Xu and Hao 2009) implemented in Xia (2013b). The four phage species colored in blue belong to Inoviridae whereas the other seven phages colored in red belong to Microviridae. The Operational Taxonomic Units (OTUs) are formed by a combination of host (the first letter of the genus name and the first four letters of the host species name), phage species name, GenBank accession number, *R* (correlation between P_U and ϕ), estimated $B_{C\to T}$, and number of tRNA genes (*N*_tRNA) in the phage genome.

the plausibility that they may have adapted to a common host that has codon usage different from that of *E. coli* and that they have invaded *E. coli* recently and have not yet had enough time to evolve codon adaptation. The shared correlation among the 12 phages with negative slopes (Table 3), summarized in the first principal component (PC1), accounts for 81% of the total variation. All 12 phages with a negative slope have P_U positively correlated with PC1 (Figure 7). Here we offer two explanations for the lack of codon adaptation in these 12 phages with empirical substantiation.

Phylogenetic inertia: If a phage has only recently invaded *E. coli*, it would have little time to evolve codon adaptation to *E. coli* translation machinery. This explanation may be applicable to phage PRD1, which has the fourth most negative slope ($b = -0.5638$, Table 3). Phage PRD1 belongs to the peculiar Tectiviridae family with members parasitizing both gram-negative and gram-positive bacteria. Phage PRD1 is the only species in the family known to parasitize a variety of gram-negative bacteria, including *Salmonella*, *Pseudomonas*, *Escherichia*, *Proteus*, *Vibrio*, *Acinetobacter*, and *Serratia* species (Bamford *et al.* 1995; Grahn *et al.* 2006). This wide host range might lead one to think that the poor codon adaptation of phage PRD1 to *E. coli* is because the phage is not *E. coli* specific. However, other lines of evidence suggest that this is not true. First, the gram-negative hosts that phage PRD1 parasitizes have similar codon usage and adaptation to any one of them or to the average of all of them

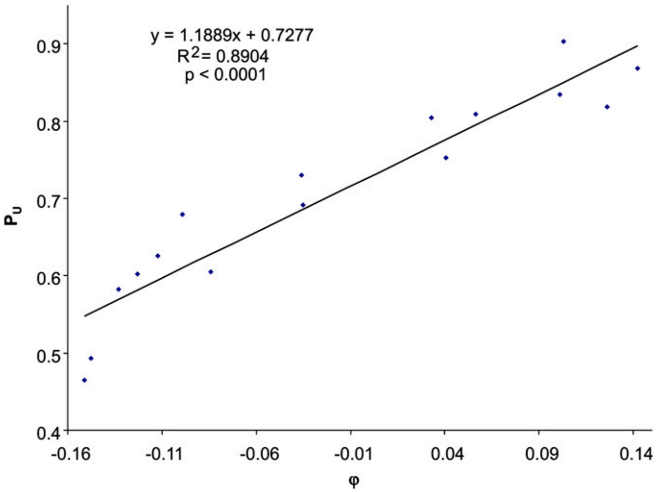


Figure 6 Relationship between P_U (the proportion of U-ending codons in Y-ending codon families) and ϕ (selection in favor of U-ending codons), based on codon usage data from *E. coli* Enterobacteria phage Phieco32 (NC_010324).

and will not lead to a negative *b* (Table 3). Second, other members of the phage family, *i.e.*, phages PR3, PR4, PR5, L17, and PR772, parasitize gram-positive bacteria. Phage PRD1 is extremely similar to its sister lineages parasitizing gram-positive bacteria. For example, there is only one amino acid difference in the coat protein between PRDl and PR4 (Bamford *et al.* 1995). It is thus quite likely that the ancestor of phage PRD1 parasitizes gram-positive bacteria. The lineage leading to phage PRD1 may have switched to gram-negative bacterial hosts only recently and still has its codon usage similar to that of its ancestral gram-positive bacterial host. In support of this, among 87 bacterial genomes covering major groups of bacterial species, the host species with codon usage most similar to that of phage PRD1 are strains in the gram-positive *Geobacillus* (NC_014206, NC_012793, NC_014650, NC_014915, and NC_013411).

Phylogenetic reconstruction with the CVTree method (Xu and Hao 2009) suggests that evolutionary history may have contributed to the differences in codon adaptation in *E. coli* dsDNA phages because codon usage captured by the intercept ($B_{C\rightarrow T}$) and the correlation (*R*) is similar among related phage species. Descendants of node A (Figure 8) belong to phage family Podoviridae, and all have high *R* 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. Another podovirus, phage Phi eco32 (NC_010324), has the highest correlation between P_U and ϕ ($r = 0.9436$). It is likely that the ancestor of podoviruses evolved good codon adaptation to an *E. coli*-like bacterial species, which was then inherited by its descendants. Because of the good codon adaptation, there is no need for the phages to carry their own tRNA genes, and all *E. coli* podoviruses studied do not encode tRNA genes in their genomes, except for phage Phi eco32 (NC_010324), which carried one putative $tRNA^{Arg}$ of uncertain function. The sequence is incorrectly annotated because the $tRNA^{Arg}$ sequence cannot be folded into a proper 7-nt anticodon loop for Arg. It also has an extraordinarily long branch when aligned and clustered with any of the *E. coli* tRNA genes, suggesting that it is unlikely to be used by *E. coli* translation machinery even if it is transcribed.

In contrast to podoviruses, a number of myoviruses have phage-encoded tRNA genes. For example, Enterobacteria phage WV8 (NC_012749) and *Erwinia* phage phiEa21-4 (NC_011811) have 19 and 23 tRNA genes, respectively (excluding one tRNA pseudogene in phage WV8). Enterobacteria phage WV8 has excellent codon adaptation, with *R* between P_U and ϕ being 0.9077. 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 genomes (Sau *et al.* 2007; Enav *et al.* 2012). Another possibility is that the Enterobacteria phage WV8 is already in the process of losing its tRNA genes because it has fewer tRNA genes than its sister lineage *Erwinia* phage phiEa21-4, which has 23 tRNA genes. Furthermore, similar to the annotated $tRNA^{Arg}$ "gene" in phage Phi eco32, annotated tRNA genes in phage WV8 are also quite different from their *E. coli* counterparts and may not be used by *E. coli* translation machinery. They may be either nonfunctional or functional in non-*E. coli* hosts.

While phage WV8 exhibits high codon adaptation to *E. coli*, other myoviruses such as those under node C have poor codon adaptation with the correlation either close to zero or negative (Figure 8, Table 3). Strong heterogeneity in codon adaptation among myoviruses is also visible among species under cluster F (Figure 8), with one myovirus having negative *R* and the rest have positive *R*. Another cluster of species with poor codon adaptation to *E. coli* (having *R* close to 0 or negative) are those under node B, made mainly of lambdoid phages. While the term "lambdoid" is never intended as a taxonomic term, the clustering of these species together suggests phylogenetic affiliations.

The lambdoid phages must have evolved in *E. coli* and *E. coli*-like hosts for a long time, and it would be weak to invoke phylogenetic inertia as an explanation for the lack of concordance of their codon usage with that of *E. coli*. This forces us to reexamine the assumptions of our model in Equations 3 and 4 in search of an alternative explanation for phages with poor codon adaptation to *E. coli*. In the two models specified in Equations 3 and 4, we have assumed a uniform mutation bias that will affect all genes and all Y-ending codon families. However, because of the asymmetric replication of the two DNA strands with associated asymmetric mutation bias, local mutation bias is often not accounted for by $P_{U.i.}$ Differences in mutation bias between the two DNA strands has been documented in organisms ranging from viruses and organelles to prokaryotic and eukaryotic genomes (Marin and Xia 2008; Xia 2012a,c).

Table 3 Results of fitting the linear regression model in Equation 3 to codon usage in dsDNA *E. coli* phages, with viral genome accession number (ACCN), the estimated intercept ($B_{C\rightarrow T}$) 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	$B_{C\rightarrow T}$	*b*	*R*	*Pa*
Phage 13a	NC_011045	0.5690	1.3782	0.8381	0.00005[b]
Phage 285P	NC_015249	0.5769	1.4737	0.8751	0.00001[b]
Phage 933W	NC_000924	0.5543	0.0089	0.0129	0.96205
Phage BP-4795	NC_004813	0.5210	−0.2133	−0.2768	0.29931
Phage CC31	NC_014662	0.7277	1.0780	0.8437	0.00004[b]
Phage cdtI	NC_009514	0.5632	−0.2310	−0.2948	0.26777
Phage EcoDS1	NC_011042	0.5477	1.4925	0.8077	0.00015[b]
Phage EPS7	NC_010583	0.7368	0.8949	0.9011	0.00000[b]
Phage HK022	NC_002166	0.4858	0.1603	0.1659	0.53920
Phage HK97	NC_002167	0.4951	0.2864	0.2699	0.31203
Phage IME08	NC_014260	0.7156	0.6011	0.6583	0.00557*c*
Phage JK06	NC_007291	0.6157	0.7818	0.5302	0.03462
Phage JS10	NC_012741	0.7206	0.6104	0.6538	0.00601[c]
Phage JS98	NC_010105	0.7200	0.6300	0.6684	0.00464[c]
Phage JSE	NC_012740	0.6931	0.6136	0.5988	0.01425[c]
Phage K1-5	NC_008152	0.6612	0.9844	0.8118	0.00013[b]
Phage K1E	NC_007637	0.6605	0.9377	0.8209	0.00010[b]
Phage K1F	NC_007456	0.6605	0.9377	0.8209	0.00010[b]
Phage lambda	NC_001416	0.4971	−0.1962	−0.1908	0.47894
Phage Min27	NC_010237	0.5504	0.0858	0.1302	0.63088
Phage Mu	NC_000929	0.4980	−0.6273	−0.5245	0.03700
Phage N15	NC_001901	0.4834	0.0208	0.0231	0.93224
Phage N4	NC_008720	0.7556	0.9110	0.8554	0.00002[b]
Phage P1	NC_005856	0.5475	0.0698	0.0885	0.74454
Phage P2	NC_001895	0.4768	−0.4703	−0.4765	0.06202
Phage P4	NC_001609	0.5065	−0.6610	−0.5315	0.03412
Phage Phi1	NC_009821	0.6905	0.5815	0.5728	0.02039[c]
Phage Phieco32	NC_010324	0.7277	1.1889	0.9436	0.00000[b]
Phage phiEcoM-GJ1	NC_010106	0.6421	1.0372	0.8690	0.00001[b]
Phage phiP27	NC_003356	0.5451	−0.3341	−0.3569	0.17473
Phage PRD1	NC_001421	0.5324	−0.5638	−0.2643	0.32262
Phage RB16	NC_014467	0.6257	1.0301	0.7718	0.00046[b]
Phage RB49	NC_005066	0.6919	0.5836	0.5790	0.01878[c]
Phage RB69	NC_004928	0.7549	0.6396	0.7414	0.00101[b]
Phage RTP	NC_007603	0.6092	0.8781	0.5433	0.02964
Phage SfV	NC_003444	0.5316	−0.1209	−0.1431	0.59705
Phage SPC35	NC_015269	0.7544	0.6200	0.8677	0.00001[b]
Phage SSL-2009a	NC_012223	0.3735	0.2555	0.3003	0.25843
Phage T1	NC_005833	0.5786	0.9154	0.5723	0.02053[c]
Phage T3	NC_003298	0.5336	1.3661	0.8663	0.00001[b]
Phage T4	NC_000866	0.7967	0.3770	0.5538	0.02603
Phage T5	NC_005859	0.7386	0.5950	0.8399	0.00005[b]
Phage T7	NC_001604	0.5641	1.3971	0.8461	0.00004[b]
Phage TLS	NC_009540	0.6377	0.5840	0.3261	0.21766
Phage vB_EcoM-VR7	NC_014792	0.6945	0.4866	0.6418	0.00736[c]
Phage VT2-Sakai	NC_000902	0.5434	0.0377	0.0517	0.84906
Phage WV8	NC_012749	0.7372	1.0841	0.9077	0.00000[b]
Phage bV_EcoS_AKFV3	NC_017969	0.7444	0.5894	0.8094	0.00015[b]
Phage D108	NC_013594	0.4994	−0.5957	−0.4791	0.06045
Phage HK639	NC_016158	0.4491	0.2538	0.2437	0.36316
Phage HK75	NC_016160	0.4839	0.3029	0.3018	0.25602
Phage phiV10	NC_007804	0.5705	0.4618	0.4478	0.08199
Phage rv5	NC_011041	0.6296	0.8548	0.6931	0.00291[b]
Phage vB_EcoM_CBA12	NC_016570	0.5988	0.3685	0.4449	0.08425
Stx1-converting bac	NC_004913	0.5512	0.0121	0.0165	0.95162
Phage BA14	NC_011040	0.5691	1.4057	0.8765	0.00001[b]
Stx2-converting phage II	NC_004914	0.5479	0.0495	0.0706	0.79503
Stx2-converting phage 1717	NC_011357	0.5011	−0.1374	−0.1724	0.52309
Stx2-converting phage 86	NC_008464	0.5681	0.0285	0.0417	0.87813
Stx2-converting phage I	NC_003525	0.5132	−0.0566	−0.0986	0.71646

[a] Significant at the 0.05 level when experimentwise error rate is controlled by the false discovery rate method.

[b] Significant with the more conservative approach of Benjamini and Yekutieli (2001).

[c] Significant with the approach of Benjamini and Hochberg (1995).

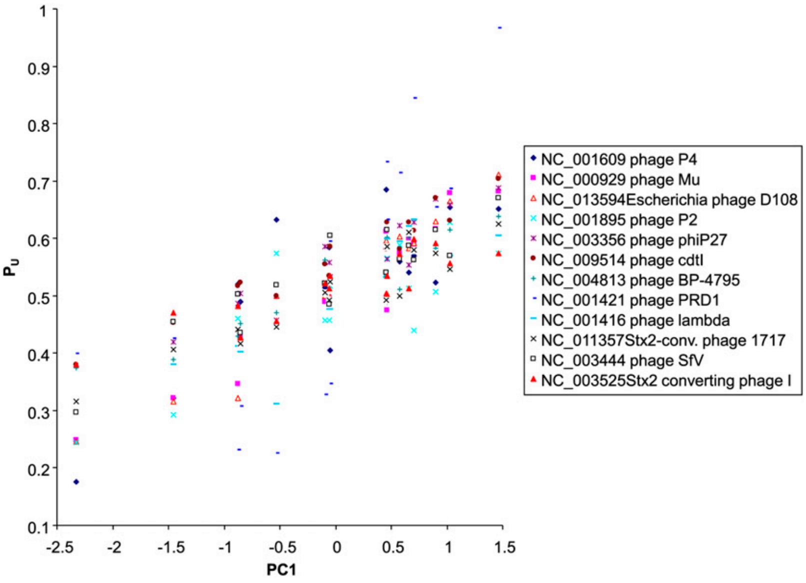


Figure 7 Phage species that do not exhibit codon adaptation to *E. coli* (negative correlation between P_U and ϕ) nevertheless exhibit codon usage similar to each other in their Y-ending codon families, with the shared correlation summarized in the first principal component (PC1) that accounts for 81% of total variation. The phage genomes are identified by their GenBank accession number and species name.

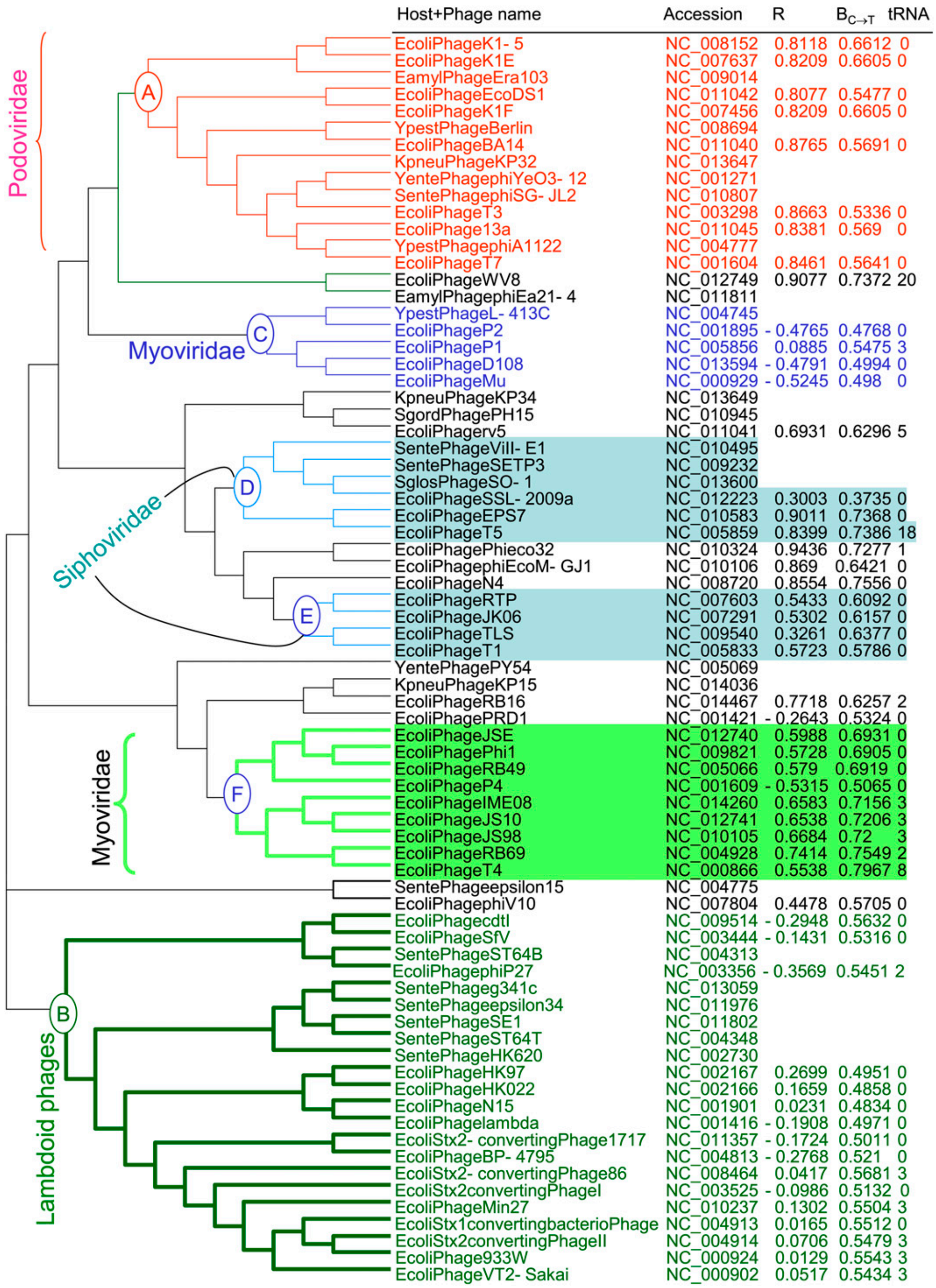
Strand asymmetry, local mutation bias, and phage codon adaptation: We found that *E. coli* dsDNA phages with negative slopes exhibit stronger local strand asymmetry than those phages with positive slopes. For example, among the 60 dsDNA *E. coli* phages, the three phages (P4, NC_001609; Mu, NC_000929; and D108, NC_013594) with the most negative *b* values (−0.6610, −0.6273, and −0.5957, respectively, Table 3) exhibit strong local TC skew, defined as $(N_T - N_C)/(N_T + N_C)$, relative to the three phage species (phage BA14, NC_011040; phage 285P, NC_015249; and phage EcoDS1, NC_011042) with the largest *b* values (Figure 9).These results suggest that codon adaptation may be difficult to achieve when one part of the genome experiences strong T-biased mutation and the other part strong C-biased mutation. Such local mutation bias is obscured if we consider only global codon frequencies over all phage genes.

How strong local strand asymmetry affects codon adaptation is not immediately obvious, so we offer an illustration here. Take, for example, the phage P4 genome (NC_001609) with 14 genes. We first need to recognize that C→T mutations often lead to not only increased U at the third codon position, but also increased U at the first and second codon positions (Figure 10). A similar response of nonsynonymous mutation rate to directional mutation pressure has also been documented in several other studies (Sueoka 1961; Lobry 2004; Urbina *et al.* 2006). The sites before genomic position 4500 (five genes) are relatively T rich and those after (nine genes) are relatively C rich (which is obvious from Figure 9). T-biased mutation reduces codons such as CCY (which is U-friendly as highly expressed *E. coli* genes strongly prefer CCU over CCC), so that CCY are found mostly at T-poor (and C-rich) regions and present mainly as CCC, which are not favored by *E. coli* translation machinery. In contrast, codons such as UUY (which is U-hostile because highly expressed *E. coli* genes strongly prefer UUC over UUU) are found mostly in T-rich regions and present mainly in the unfavored UUU form. Thus, the T-rich region features many unfavored UUU codons and C-rich regions feature many unfavored CCC codons, leading to poor codon adaptation.

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}). Take the six dsDNA phage species in Figure 9, for example. The I_{SA} value would be much greater for the three species with negative slopes than for the three species with positive slopes. The *R* value is highly significantly and negatively correlated with I_{SA} for ssDNA phages ($P = 0.0008$, Figure 11). 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 Equations 3 and 4 to phage genomic segments with relatively homogenous mutation bias.

The relationship between *R* and I_{SA} in Figure 11 offers an explanation for the outlying point in Figure 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) among ssDNA phages and in this new light is expected to be associated with a low *R* value (Figure 11). In short, the mutation effect (on the *x*-axis) for Enterobacteria phage If1 is underestimated in Figure 4, which does not take local strand asymmetry into consideration. When local strand asymmetry is accounted for (Figure 11), the point is shifted rightward along the *x*-axis to its proper location.

We may conclude that selection on Y-ending codons, represented by ϕ, detectable in dsDNA phages as the mean



Host+Phage name	Accession	R	$B_{C\rightarrow T}$	tRNA
EcoliPhageK1- 5	NC_008152	0.8118	0.6612	0
EcoliPhageK1E	NC_007637	0.8209	0.6605	0
EamylPhageEra103	NC_009014			
EcoliPhageEcoDS1	NC_011042	0.8077	0.5477	0
EcoliPhageK1F	NC_007456	0.8209	0.6605	0
YpestPhageBerlin	NC_008694			
EcoliPhageBA14	NC_011040	0.8765	0.5691	0
KpneuPhageKP32	NC_013647			
YentePhagephiYeO3- 12	NC_001271			
SentePhagephiSG- JL2	NC_010807			
EcoliPhageT3	NC_003298	0.8663	0.5336	0
EcoliPhage13a	NC_011045	0.8381	0.569	0
YpestPhagephiA1122	NC_004777			
EcoliPhageT7	NC_001604	0.8461	0.5641	0
EcoliPhageWV8	NC_012749	0.9077	0.7372	20
EamylPhagephiEa21- 4	NC_011811			
YpestPhageL- 413C	NC_004745			
EcoliPhageP2	NC_001895	- 0.4765	0.4768	0
EcoliPhageP1	NC_005856	0.0885	0.5475	3
EcoliPhageD108	NC_013594	- 0.4791	0.4994	0
EcoliPhageMu	NC_000929	- 0.5245	0.498	0
KpneuPhageKP34	NC_013649			
SgordPhagePH15	NC_010945			
EcoliPhagerv5	NC_011041	0.6931	0.6296	5
SentePhageVilI- E1	NC_010495			
SentePhageSETP3	NC_009232			
SglosPhageSO- 1	NC_013600			
EcoliPhageSSL- 2009a	NC_012223	0.3003	0.3735	0
EcoliPhageEPS7	NC_010583	0.9011	0.7368	0
EcoliPhageT5	NC_005859	0.8399	0.7386	18
EcoliPhagePhieco32	NC_010324	0.9436	0.7277	1
EcoliPhagephiEcoM- GJ1	NC_010106	0.869	0.6421	0
EcoliPhageN4	NC_008720	0.8554	0.7556	0
EcoliPhageRTP	NC_007603	0.5433	0.6092	0
EcoliPhageJK06	NC_007291	0.5302	0.6157	0
EcoliPhageTLS	NC_009540	0.3261	0.6377	0
EcoliPhageT1	NC_005833	0.5723	0.5786	0
YentePhagePY54	NC_005069			
KpneuPhageKP15	NC_014036			
EcoliPhageRB16	NC_014467	0.7718	0.6257	2
EcoliPhagePRD1	NC_001421	- 0.2643	0.5324	0
EcoliPhageJSE	NC_012740	0.5988	0.6931	0
EcoliPhagePhi1	NC_009821	0.5728	0.6905	0
EcoliPhageRB49	NC_005066	0.579	0.6919	0
EcoliPhageP4	NC_001609	- 0.5315	0.5065	0
EcoliPhageIME08	NC_014260	0.6583	0.7156	3
EcoliPhageJS10	NC_012741	0.6538	0.7206	3
EcoliPhageJS98	NC_010105	0.6684	0.72	3
EcoliPhageRB69	NC_004928	0.7414	0.7549	2
EcoliPhageT4	NC_000866	0.5538	0.7967	8
SentePhageepsilon15	NC_004775			
EcoliPhagephiV10	NC_007804	0.4478	0.5705	0
EcoliPhagecdtI	NC_009514	- 0.2948	0.5632	0
EcoliPhageSfV	NC_003444	- 0.1431	0.5316	0
SentePhageST64B	NC_004313			
EcoliPhagephiP27	NC_003356	- 0.3569	0.5451	2
SentePhageg341c	NC_013059			
SentePhageepsilon34	NC_011976			
SentePhageSE1	NC_011802			
SentePhageST64T	NC_004348			
SentePhageHK620	NC_002730			
EcoliPhageHK97	NC_002167	0.2699	0.4951	0
EcoliPhageHK022	NC_002166	0.1659	0.4858	0
EcoliPhageN15	NC_001901	0.0231	0.4834	0
EcoliPhagelambda	NC_001416	- 0.1908	0.4971	0
EcoliStx2- convertingPhage1717	NC_011357	- 0.1724	0.5011	0
EcoliPhageBP- 4795	NC_004813	- 0.2768	0.521	0
EcoliStx2- convertingPhage86	NC_008464	0.0417	0.5681	3
EcoliStx2convertingPhageI	NC_003525	- 0.0986	0.5132	0
EcoliPhageMin27	NC_010237	0.1302	0.5504	3
EcoliStx1convertingbacterioPhage	NC_004913	0.0165	0.5512	0
EcoliStx2convertingPhageII	NC_004914	0.0706	0.5479	3
EcoliPhage933W	NC_000924	0.0129	0.5543	3
EcoliPhageVT2- Sakai	NC_000902	0.0517	0.5434	3

Figure 8 Phylogenetic tree of dsDNA phages reconstructed by using the CVTree method with k (peptide length) = 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 species name), phage species name, GenBank accession number, R (correlation between P_U and ϕ), estimated $B_{C\rightarrow T}$, and number of tRNA genes (tRNA) in the phage genome.

of the b values (Table 3) is significantly >0 (mean b = 0.4622, SE = 0.07436, t = 6.2156, d.f. = 59, $P < 0.0001$). When the false discovery method is used to control for type I error rate, about half of the b values are statistically significant at the 0.05 level (Table 3).

Other factors that may contribute to phage codon usage

One factor that may contribute to phage codon usage bias is phage-encoded tRNA genes. Note that *E. coli* dsDNA phages other than those in clusters A, B, and C (Figure 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

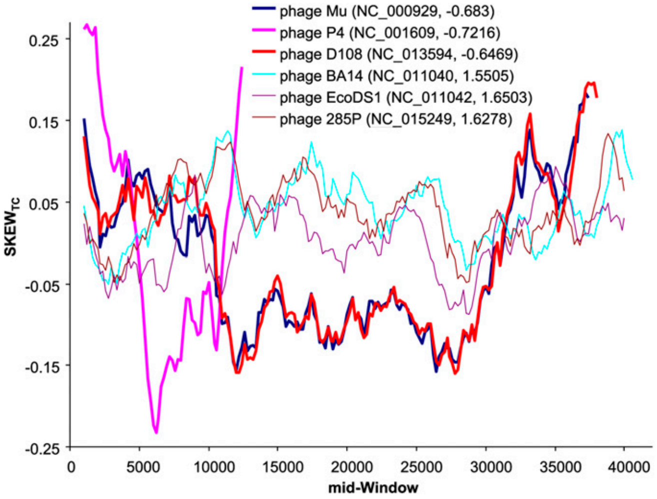


Figure 9 TC skew, defined as $(N_T - N_C)/(N_T + N_C)$, for three phages with the most negative slopes and three phages with the most positive slopes, plotted over sliding windows (window width = 2000 nt and step size = 200). Phage P4 is much shorter than the others and requires coinfection by phage P2 to complete its lytic cycle. The key shows the phage name with the GenBank accession number and the slope from regression of P_U on ϕ.

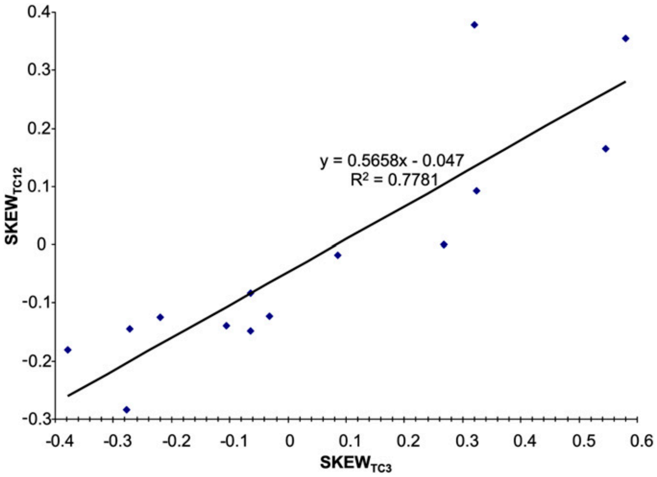


Figure 10 TC skew at the first and second codon positions (SKEW$_{TC12}$) in the 14 genes in phage P4 increases with TC skew at the third codon position (SKEW$_{TC3}$).

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 favoring 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 that are typically translated by two types of tRNAs, one with a wobble U consistently preferring A-ending codons and the other with 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 bioinformatically.

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 a high-temperature environment and have relatively high genomic GC. Their phages are also expected to have high GC to maintain genome stability at high temperature (Xia and Yuen 2005). Different host species may have the 4 nt 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 and Palidwor 2005; Marin and 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.

Dinucleotide frequencies are often used to explore the presence of site-dependent mutation patterns. Some bacterial species such as *Mycoplasma pulmonis* carry CpG-specific methyltransferase and exhibit strong CpG deficiency (Xia 2003) that could lead to context-dependent codon usage; *e.g.*, C-ending codons are particularly rare if the next codon is GNN (where N stands for any nucleotide). However, neither *E. coli* nor any of its phages carries CpG-specific methyltransferase genes. The ratio $P_{CpG}/(P_C P_G)$ is close to 1 for both *E. coli* and its phages.

Given nucleotide frequencies P_i (where i = A, C, G, or T), the observed dinucleotide frequencies P_{ij}, assuming random association, are expected to be $P_i P_j$ (where i, j = A, C, G, or T). The deviation of the observed frequency from the expected frequency may be expressed as $D_{ij} = (P_{ij} - P_i P_j)/P_i P_j$. A dinucleotide is in surplus if $D_{ij} > 0$ and in deficiency if $D_{ij} < 0$. The 11 ssDNA phages share similar dinucleotide frequencies, all with D_{AA} being the highest and D_{AG} the lowest. However, the surplus of AA dinucleotides is largely explainable by the usage of AAN codons, especially AAA codons, which is far more than expected. If we designate total number of codons in the 11 ssDNA phages as N_T, then the expected number of AAA codons (designated E_{AAA}) is 404.7 ($= N_T \times P_A^3$) whereas the observed AAA codon (designated O_{AAA}) is 930. O_{AAG}, O_{AAC}, and O_{AAU} codons are also greater than E_{AAG}, E_{AAC} and E_{AAU}, although not as dramatic as AAA codons. In contrast, the deficiency of AG

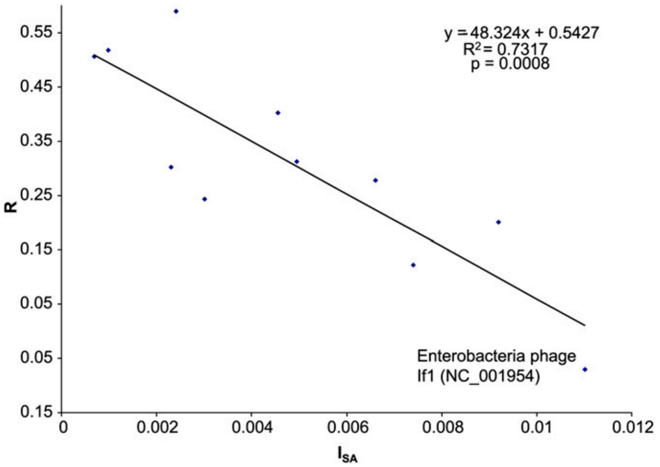


Figure 11 The effect of selection on Y-ending codons, measured by the correlation (R) between P_U and ϕ, decreases with the degree of strand asymmetry, measured by the index of strand asymmetry (I_{SA}, which is the variance of the window-specific TC skew values).

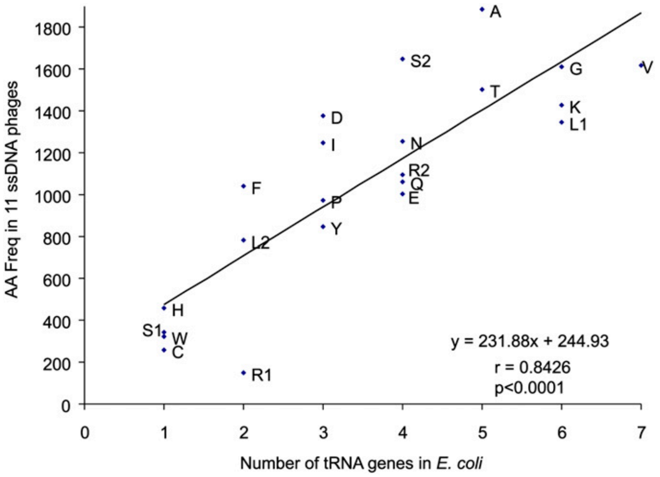


Figure 12 Frequency of amino acids increases with the number of tRNA genes in *E. coli* K12 MG1655. Met tRNAs are not included because Met is translated by both initiator and elongator tRNAs while all other amino acids are by elongator tRNAs only.

dinucleotides can be attributed to O_{AGA}, O_{AGG}, O_{AGC}, and O_{AGT} (= 104, 44, 163, and 179, respectively) all being smaller than E_{AGA}, E_{AGG}, E_{AGC}, and E_{AGT} (= 351.5, 305.3, 306.9, and 430.6, respectively).

One may ask why O_{AAN} is far greater than E_{AAN} whereas O_{AGN} is far smaller than E_{AGN}. One simple explanation invokes tRNA abundance (Xia 1998), which is well predicted by tRNA copy number (Percudani *et al.* 1997). In unicellular organisms such as *E. coli*, *Salmonella typhimurium*, and *Saccharomyces cerevisiae*, the frequency of an amino acid increases with the abundance of tRNA carrying the amino acid. As ssDNA phages do not carry their own tRNA genes and therefore depend entirely on the host tRNA pool to translate phage genes, we expect the frequency of amino acids to increase with the number of their cognate tRNA genes in the *E. coli* genome. This expectation is supported by a strong positive correlation between the frequency of an amino acid and the number of gene copies of the tRNA carrying the amino acid ($r = 0.8426$, $P < 0.0001$, Figure 12). This finding explains why there is a surplus of AA dinucleotides because there are six tRNA genes decoding AAA and AAG codons, leading to relative overuse of AAR codons. In contrast, there are only two tRNA genes for AGR codons (coding Arg1) and only one tRNA for the AGY codons (coding for Ser1), which explains the relatively rarity of AGR and AGY codons as well as the deficiency of AG dinucleotides mentioned above.

We did not incorporate translation initiation into our model in this article. However, selection on codon adaptation is present only for mRNAs with efficient initiation; *i.e.*, one should expect little selection for codon adaptation in genes whose mRNAs have poor translation initiation efficiency. The availability of ribosomal loading data and their analysis (Xia *et al.* 2011) may eventually lead to an index of translation initiation to facilitate more vigorous studies of codon adaptation.

We should mention that, although selection on codon usage appears more detectable in dsDNA phages than in ssDNA phages (Table 2 and Table 3), this does not mean that dsDNA phage mRNA will necessarily have better translation efficiency than ssDNA phage mRNA. A previous study suggested that dsDNA phage coat protein genes have higher CAI (Sharp and Li 1987), which has been known to be a good measure of translation efficiency (Comeron and Aguade 1996; Duret and Mouchiroud 1999; Coghlan and Wolfe 2000), than ssDNA phage genes. However, there is no coat gene that is homologous between dsDNA and ssDNA phages, so the comparison may be between apples and oranges. If all genes are included, then the difference is minimal (mean CAI = 0.4768 for dsDNA phages and 0.4743 for ssDNA phages, excluding the 22 dsDNA phages with phage-encoded tRNA genes) and not statistically significant. The mean CAI value for the 22 dsDNA phages with phage-encoded tRNA genes is even lower, but that is because the phage-encoded tRNAs may allow the phage to use codons frequently used in the phage but rare in the host (*i.e.*, the phage tRNAs reduce the need for the phage to evolve codon usage similar to that of the host). Some ssDNA phages with increasing C→T mutation bias appear to increase the usage of codons in the "U-friendly" codon families, thereby achieving CAI values almost as large as those of dsDNA phages.

In summary, 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 ($B_{C\rightarrow T}$) and to evaluate the relative effect of the mutation bias and host translation machinery on phage codon usage. C→T mutations occur more frequently in ssDNA phages than in dsDNA phages and affect not only synonymous codon usage, but also nonsynonymous substitutions, especially in ssDNA phages. dsDNA phages exhibit better codon adaptation to host translation machinery than ssDNA phages, but much of the variation in codon usage may be attributed to phylogenetic inertia. Strand asymmetry strongly influences the efficiency of selection on codon adaptation and needs to be taken into account when studying codon adaptation.

Acknowledgments

We thank S. Aris-Brosou, N. Corradi, M. Gilchrist, A. Golshani, and J. Wang for discussion and comments. Two anonymous reviews and comments from J. Lawrence resulted in significant improvement of the manuscript. This study is supported by a Natural Science and Engineering Research Council of Canada discovery grant (to X.X.).

Literature Cited

Abedon, S. T., S. J. Kuhl, B. G. Blasdel, and E. M. Kutter, 2011 Phage treatment of human infections. Bacteriophage 1: 66–85.

Azeredo, J., and I. W. Sutherland, 2008 The use of phages for the removal of infectious biofilms. Curr. Pharm. Biotechnol. 9: 261–266.

Bamford, D. H., J. Caldentey, and J. K. Bamford, 1995 Bacteriophage PRD1: a broad host range DSDNA tectivirus with an internal membrane. Adv. Virus Res. 45: 281–319.

Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57: 289–300.

Benjamini, Y., and D. Yekutieli, 2001 The control of the false discovery rate in multiple hypothesis testing under dependency. Ann. Stat. 29: 1165–1188.

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.

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

Carullo, M., and X. Xia, 2008 An extensive study of mutation and selection on the wobble nucleotide in tRNA anticodons in fungal mitochondrial genomes. J. Mol. Evol. 66: 484–493.

Coghlan, A., and K. H. Wolfe, 2000 Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. Yeast 16: 1131–1145.

Comeron, J. M., and M. Aguade, 1996 Synonymous substitutions in the Xdh gene of *Drosophila*: heterogeneous distribution along the coding region. Genetics 144: 1053–1062.

Duffy, S., and E. C. 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., and E. C. 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.

Duncan, B. K., and J. H. Miller, 1980 Mutagenic deamination of cytosine residues in DNA. Nature 287: 560–561.

Duret, L., and D. Mouchiroud, 1999 Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. Proc. Natl. Acad. Sci. USA 96: 4482–4487.

Edwards, R. A., and F. Rohwer, 2005 Viral metagenomics. Nat. Rev. Microbiol. 3: 504–510.

Enav, H., O. Beja, and 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. Am. Nat. 125: 1–15.

Felsenstein, J., 2004 *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

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

Gouy, M., 1987 Codon contexts in enterobacterial and coliphage genes. Mol. Biol. Evol. 4: 426–444.

Gouy, M., and C. Gautier, 1982 Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 10: 7055–7064.

Grahn, A. M., S. J. Butcher, J. K. H. Bamford, and D. H. Bamford, 2006 PRD1: dissecting the genome, structure and entry, pp. 176–185 in *The Bacteriophages*, edited by R. Calendar. Oxford University Press, Oxford.

Grosjean, H., D. Sankoff, W. M. Jou, W. Fiers, and R. J. 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, and B. Seed, 1996 Codon usage limitation in the expression of HIV-1 envelope glycoprotein. Curr. Biol. 6: 315–324.

Hernan, R. A., H. L. Hui, M. E. Andracki, R. W. Noble, S. G. Sligar *et al.*, 1992 Human hemoglobin expression in *Escherichia coli*: importance of optimal codon usage. Biochemistry 31: 8619–8628.

Higgs, P. G., and 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.

Ikemura, T., 1981 Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J. Mol. Biol. 146: 1–21.

Ikemura, T., 1992 Correlation between codon usage and tRNA content in microorganisms, pp. 87–111 in *Transfer RNA in Protein Synthesis*, edited by D. L. Hatfield, B. J. Lee, and R. M. Pirtle. CRC Press, Boca Raton, FL.

Jia, W., and P. G. Higgs, 2008 Codon usage in mitochondrial genomes: distinguishing context-dependent mutation from translational selection. Mol. Biol. Evol. 25: 339–351.

Kleber-Janke, T., and W. M. Becker, 2000 Use of modified BL21 (DE3) *Escherichia coli* cells for high-level expression of recombinant peanut allergens affected by poor codon usage. Protein Expr. Purif. 19: 419–424.

Koresawa, Y., S. Miyagawa, M. Ikawa, K. Matsunami, M. Yamada *et al.*, 2000 Synthesis of a new Cre recombinase gene based on optimal codon usage for mammalian systems. J. Biochem. 127: 367–372.

Kreutzer, D. A., and J. M. Essigmann, 1998 Oxidized, deaminated cytosines are a source of C → T transitions in vivo. Proc. Natl. Acad. Sci. USA 95: 3578–3582.

Kunisawa, T., S. Kanaya, and E. Kutter, 1998 Comparison of synonymous codon distribution patterns of bacteriophage and host genomes. DNA Res. 5: 319–326.

Lindahl, T., 1993 Instability and decay of the primary structure of DNA. Nature 362: 709–715.

Lobry, J. R., 2004 Life history traits and genome structure: aerobiosis and G+C content in bacteria. Lect. Notes Comput. Sci. 3039: 679–686.

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

Marin, A., and 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.

Mitra, S. K., F. Lustig, B. Akesson, and U. Lagerkvist, 1977 Codon-anticodon recognition in the valine codon family. J. Biol. Chem. 252: 471–478.

Mitra, S. K., F. Lustig, B. Akesson, T. Axberg, P. Elias *et al.*, 1979 Relative efficiency of anticodons in reading the valine codons during protein synthesis in vitro. J. Biol. Chem. 254: 6397–6401.

Nasvall, S. J., P. Chen, and G. R. Bjork, 2007 The wobble hypothesis revisited: uridine-5-oxyacetic acid is critical for reading of G-ending codons. RNA 13: 2151–2164.

Ngumbela, K. C., K. P. Ryan, R. Sivamurthy, M. A. Brockman, R. T. Gandhi *et al.*, 2008 Quantitative effect of suboptimal codon usage on translational efficiency of mRNA encoding HIV-1 *gag* in intact T cells. PLoS ONE 3: e2356.

Palidwor, G. A., T. J. Perkins, and 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 *et al.*, 2013 Vaccinia and influenza A viruses select rather than adjust tRNAs to optimize translation. Nucleic Acids Res. 41: 1914–1921.

Percudani, R., A. Pavesi, and S. Ottonello, 1997 Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. J. Mol. Biol. 268: 322–330.

Ranjan, A., A. S. Vidyarthi, and R. Poddar, 2007 Evaluation of codon bias perspectives in phage therapy of *Mycobacterium tuberculosis* by multivariate analysis. In Silico Biol. 7: 423–431.

Rice, P., I. Longden, and A. Bleasby, 2000 EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16: 276–277.

Rohwer, F., and R. Edwards, 2002 The Phage Proteomic Tree: a genome-based taxonomy for phage. J. Bacteriol. 184: 4529–4535.

Sahu, K., S. K. Gupta, S. Sau, and T. C. Ghosh, 2005 Comparative analysis of the base composition and codon usages in fourteen mycobacteriophage genomes. J. Biomol. Struct. Dyn. 23: 63–71.

SAS Institute, 1994 *SAS/STAT User's Guide*, Vol. 2. GLM-VARCOMP, Cary, NC: SAS Institute Inc.

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., S. K. Gupta, S. Sau, and T. C. Ghosh, 2005 Synonymous codon usage bias in 16 *Staphylococcus aureus* phages: implication in phage therapy. Virus Res. 113: 123–131.

Sau, K., S. K. Gupta, S. Sau, S. C. Mandal, and T. C. 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., A. N. Brooks, and T. M. Lowe, 2005 The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 33: W686–W689.

Shackelton, L. A., and E. C. Holmes, 2006 Phylogenetic evidence for the rapid evolution of human B19 erythrovirus. J. Virol. 80: 3666–3669.

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

Sharp, P. M., and W. H. Li, 1987 The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15: 1281–1295.

Sueoka, N., 1961 Correlation between base composition of deoxyribonucleic acid and amino acid composition of proteins. Proc. Natl. Acad. Sci. USA 47: 1141–1149.

Sun, X. Y., and Q. Yang, and X. Xia, 2013 An improved implementation of effective number of codons (N_c). Mol. Biol. Evol. 30: 191–196.

Umemura, T., Y. Tanaka, K. Kiyosawa, H. J. Alter, and J. W. 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, and P. G. 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. J. Mol. Evol. 62: 340–361.

van Weringh, A., M. Ragonnet-Cronin, E. Pranckeviciene, M. Pavon-Eternod, L. Kleiman *et al.*, 2011 HIV-1 modulates the tRNA pool to improve translation efficiency. Mol. Biol. Evol. 28: 1827–1834.

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., 2003 DNA methylation and mycoplasma genomes. J. Mol. Evol. 57: S21–S28.

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. Evol. Bioinform. 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 Position weight matrix, Gibbs sampler, and the associated significance tests in motif characterization and prediction. Scientifica 2012: 917540.

Xia, X., 2012c Rapid evolution of animal mitochondria, pp. 73–82 in *Evolution in the Fast Lane: Rapidly Evolving Genes and Genetic Systems*, edited by R. S. Singh, J. Xu, and R. J. Kulathinal. Oxford University Press, Oxford.

Xia, X., 2013a *Comparative Genomics*. Springer Heidelberg.

Xia, X., 2013b DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. Mol. Biol. Evol. 30: 1720–1728.

Xia, X., and G. Palidwor, 2005 Genomic adaptation to acidic environment: evidence from *Helicobacter pylori*. Am. Nat. 166: 776–784.

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

Xia, X., H. Huang, M. Carullo, E. Betran, and E. N. Moriyama, 2007 Conflict between translation initiation and elongation in vertebrate mitochondrial genomes. PLoS ONE 2: e227.

Xia, X., V. MacKay, X. Yao, J. Wu, F. Miura *et al.*, 2011 Translation initiation: a regulatory role for poly(A) tracts in front of the AUG codon in *Saccharomyces cerevisiae*. Genetics 189: 469–478.

Xu, Z., and B. Hao, 2009 CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. Nucleic Acids Res. 37: W174–W178.

Communicating editor: J. Lawrence