DNA Research, 2017, 24(4), 333–342
doi: 10.1093/dnares/dsx005
Advance Access Publication Date: 28 February 2017
Full Paper

OXFORD

Full Paper

Evidence of translation efficiency adaptation of the coding regions of the bacteriophage lambda

Eli Goz[1,2,†], Oriah Mioduser[1,†], Alon Diament[1], and Tamir Tuller[1,2,3,*]

[1]Department of Biomedical Engineering, Tel-Aviv University, Ramat Aviv 69978, Israel, [2]SynVaccine Ltd Ramat Hachayal, Tel Aviv 6971039, Israel, and [3]Sagol School of Neuroscience, Tel-Aviv University, Ramat Aviv 69978, Israel

*To whom correspondence should be addressed. Tel. 972 - 3-6405839. Fax. 972 - 3-6407308. Email: tamirtul@post.tau.ac.il

[†]These authors contributed equally to this study.

Edited by Prof. Kenta Nakai

Received 6 October 2016; Editorial decision 23 January 2017; Accepted 1 February 2017

Abstract

Deciphering the way gene expression regulatory aspects are encoded in viral genomes is a challenging mission with ramifications related to all biomedical disciplines. Here, we aimed to understand how the evolution shapes the bacteriophage lambda genes by performing a high resolution analysis of ribosomal profiling data and gene expression related synonymous/silent information encoded in bacteriophage coding regions.

We demonstrated evidence of selection for distinct compositions of synonymous codons in early and late viral genes related to the adaptation of translation efficiency to different bacteriophage developmental stages. Specifically, we showed that evolution of viral coding regions is driven, among others, by selection for codons with higher decoding rates; during the initial/progressive stages of infection the decoding rates in early/late genes were found to be superior to those in late/early genes, respectively. Moreover, we argued that selection for translation efficiency could be partially explained by adaptation to *Escherichia coli* tRNA pool and the fact that it can change during the bacteriophage life cycle.

An analysis of additional aspects related to the expression of viral genes, such as mRNA folding and more complex/longer regulatory signals in the coding regions, is also reported. The reported conclusions are likely to be relevant also to additional viruses.

Key words: decoding rates, bacteriophage genome, bacteriophage fitness, coding regions evolution, ribosomes and tRNAs

1. Introduction

Because of a major advantage of being a diverse group of easily manipulated viruses, bacteriophages have various potential uses both in fundamental research and in various biotechnological and biomedical applications. For example, they are used as vehicles for vaccines (both DNA and protein), for the detection of pathogenic bacterial strains, and as a display system for many proteins and antibodies.[1–4] Furthermore, phages were also suggested to be used as biocontrol agents in agriculture and petroleum industry, and as alternatives to antibiotics in the case of antibiotic resistant bacterial strains. In addition, they often serve as model organisms in molecular evolution studies.[1–4] Therefore, understanding the way the viral fitness is encoded in the genetic material of bacteriophages (or other type of viruses) is an important and challenging mission that may potentially contribute to all biomedical disciplines.[1–4]

Deciphering the regulatory information encoded in the genomes of phages or other viruses, and the relation between the nucleotides composition of the coding regions and the viral fitness is a very

challenging mission, which was tackled, in the recent years, by various researchers.[5–13] Among others, it was suggested that ribosome profiling, which enables the large *in vivo* genome wide monitoring of ribosome state at a resolution of single nucleotide, is a very useful tool for deciphering the coding complexity of viral (and other organisms) genomes. Specifically, it was shown that ribosome profiling enables detecting novel (possibly very short) coding regions and estimate the translation status of various open reading frames.[5–9]

Here, we focus on the Bacteriophage lambda which is a well-known and studied member of the Siphoviridae family of double-stranded DNA viruses in the Caudovirales order (also known as 'tailed bacteriophages' due to their characteristic form). During its lifecycle this phage either resides within the genome of its *Escherichia coli* host through lysogeny or enters into a lytic phase (which lasts ∼25 min) during which it produces progeny viral particles, and kills and lyses the cell (see for example[14,15]). The genome size of the bacteriophage lambda is ∼50 kb nt and includes 66 known genes that were analysed in this study. These genes can be divided into two groups, 'early' and 'late' according to the stages in the lytic phase when their expression is dominant.[9]

The specific aim of this study is at exploring the way in which translation efficiency related information is encoded on a synonymous level in the coding regions of genes that are expressed during different bacteriophage lambda development stages (i.e. 0–20 min after the beginning of the lytic phase). Some previous studies shed some light on specific elements related to this topic. For example, a recent study by Liu et al.[9] employed ribosome profiling to study the progress of bacteriophage lambda gene expression during phage development and showed that the known genes are expressed in a predictable fashion; in addition many previously unknown potential open reading frames were detected. Other studies focused on different aspects of viral translation (and lifecycle in general) regulation that has a significant effect on shaping viral genomes, such as: secondary structures within viral transcripts (mainly in untranslated regions[16,17] but also within coding regions[18–20]) and evolutionary pressure on synonymous codons usage bias.[21–25]

However, to the best of our knowledge, this is the first study aiming to perform a comprehensive, large scale analysis of different types of genomic synonymous information related to regulation of translation efficiency in all coding regions of a bacteriophage lambda, during different stages of its development. In particular, basing on the analysis of ribosome profiling data,[9] we demonstrated, for the first time, the condition specific adaptation of the bacteriophage codons in early/late genes to the intra-cellular *E. coli* environment during the different stages of phage development.

2. Material and methods

2.1. Data

Ribosome profiling was applied to the process of lytic growth of Bacteriophage lambda by Liu et al.[9] They chose temperature induction of the classic cI857 repressor mutant of Bacteriophage lambda in a lysogen of *E. coli* MG1655 to synchronize the lytic process, sampling the lysogen and control non-lysogen both before and 2, 5, 10, and 20 min after shifting the temperature from 32 °C to 42 °C. Transcript sequences were obtained from EnsEMBL for *E. coli* (K-12 MG1655 release 121, accessed 28/07/15) and from NCBI for the lambda phage (J02459, accessed 07/12/15). There were 4,319 genes of *E. coli* and 66 genes of lambda phage in the obtained sequences.

2.1.1. Ribo-seq reads mapping

Ribosome footprint sequences were obtained from[9] (GSE47509, induction 0–20 min). We trimmed the poly-A adaptors from the reads using Cutadapt[26] (version 1.8.3), and utilized Bowtie[27] (version 1.1.1) to map them to the *E. coli*-lambda transcriptome. The location of the A-site was approximated by an 11-nt shift from the 5′ end of the aligned read. This shift maximized the correlation between MTDR (described below) and the observed read densities per *E. coli* gene. Further details related to the ribo-seq processing appear in the supplementary methods.

2.2. Randomization models

We considered two randomization models: (i) To preserve both *the amino acids order and content* and the frequencies distribution of 16 possible pairs of adjacent nucleotides (*dinucleotides*), a model based on a multivariate Boltzmann sampling scheme was used.[28] This model was initially introduced in the context of enumerative combinatorics and was used by us before for studying other viruses.[16,17] We used the original source code which can be found in http://csb.cs.mcgill.ca/sparcs (7 February 2017, date last accessed). (ii) To preserve both *the amino acids order and content* and the codon usage bias, synonymous positions in codon sequences were randomly permuted.

2.2.1. tRNA adaptation index (*tAI*).[29]

Quantifies the adaptation of the codons of a coding region to the tRNA pool. Technical details regarding this measure appear in the supplementary.

2.2.2. Ribosome profiling data normalization

We began this analysis by reconstructing ribosome profiles for *E. coli* and Bacteriophage Lambda expressed genes and performing normalization described in the supplementary data. The normalization enables measuring the relative time a ribosome spends translating each codon in a specific gene relative to other codons in it, whilst considering the total number of codons in the gene, resulting in its normalized footprint count (*NFC*)[30]:

$$NFC_j = \frac{RC_j}{\left(\frac{1}{J-40}\right)\left(RC_{21} + RC_{22} + \ldots + RC_{J-20}\right)},$$

$$j = 21 \ldots J - 20,$$

where J is the number of codons in the gene and j is the index of a codon.

We generate histogram of *NFC* for each codon. Each *NFC* distribution describes the probability (Pi) (y-axis) of observing each of the codon's *NFC* values (x-axis) in the ORFs of the analysed organism.

2.3. Codon typical decoding rate (TDR)

To estimate the typical decoding time of each codon based on *NFC* distributions, we used a novel statistical model,[17] which takes into consideration the skewed nature of the *NFC* distribution. The aim is to describe the *NFC* histogram of each codon as an output of a random variable which is a sum of two random variables: a normal and an exponential variable. Thus, the distribution of this new random variable includes three parameters, and is called *EMG* distribution.[30] In this model, the typical codon decoding time was described by the

normal distribution with two parameters: mean (μ) and standard deviation (σ); the μ parameter represents the location of the mean of the theoretical Gaussian component that should be obtained if there are no phenomena such as pauses/biases/ribosomal traffic jams; σ represents the width of the Gaussian component. The exponential distribution has one parameter λ which represents the skewness of the *NFC* distribution due to reasons such as ribosomal jamming caused by codons with different decoding times, extreme pauses, incomplete halting of the ribosomes and biases in the experiment. The *EMG* is defined as follows:

$$f(x;\mu,\sigma,\lambda)=\frac{\lambda}{2}e^{\frac{\lambda}{2}\left(2\mu+\lambda\sigma^2-2x\right)}erfc\left(\frac{\mu+\lambda\sigma^2-x}{\sqrt{2}\sigma}\right),$$

$$erfc(x)=1-erf(x)=\int_x^{\infty}e^{-t^2}dt.$$

We used the maximum likelihood criterion to estimate these three parameters for each codon based on *E. coli* ribosome profiling data by fitting the suggested model to the *NFC* distribution. $1/\mu$ was defined to be the *TDR* of each codon.

In order to optimize the *TDR* to *E. coli*'s read counts in every time condition; we removed outliers from the *NFC* distribution of each codon in the following way: for every codon (in every time condition), and for each *NFCi* point related to the codon, we calculated the *Pi* to see value larger or equal to *NFCi* based on the pdf fitted to the codon (*EMG* distribution). Let *Ni* denote the number of measurements of the codon *NFC* based on the data; we removed points in which the result of $p_i{*}N_i$ was lower than 0.001.

2.4. Synonymous codons usage analysis

Synonymous codons composition of a coding sequence was represented by a 61-dimensional vector of relative synonymous codons frequencies (*RSCF*) of each one of 61 coding codons (stop codons were excluded):

$$RSCF=(RSCF[1],\ldots,RSCF[61]),$$

$$RSCF[i]=\frac{q_i}{\sum_{j\in syn[i]}q_j},\quad\sum_{j\in syn[i]}RSCF[j]=1,$$

where q_i is the number of appearances of codon i in a sequence, $syn[i]$ is a subset of indexes in *RSCF* pointing at codons synonymous to codon i. Therefore, each of the 61 coding codons was assigned a number between 0 and 1 according to its frequency relatively to the other codons coding for the same among acid.

Clustering analysis was performed on *RSCF* vectors of all coding sequences. In order to exclude biases due to a possible absence of specific amino acids in some of the sequences (missing amino acids), the relative synonymous frequency of a codon corresponding to a missing amino acid was set to the average relative synonymous frequency of this codon over all sequences in which at least one such amino acid is present.

Each viral sequence was assigned a group label corresponding to its temporary expression stage (early/late) (according to the classification known in the literature). The tendency of sequences to cluster according to the codons usage in two different clusters corresponding to their temporary expression stages (early/late) was measured using the Davies–Bouldin score (*DBS*). This score is based on a ratio of within-cluster and between-cluster distances and is defined as:

$$DB=\frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\{D_{ij}\},D_{ij}=\frac{\bar{d}_i+\bar{d}_j}{d_{ij}},$$

where k is the number of evaluated clusters, D_{ij} is the within-to-between cluster distance ratio for the ith and jth clusters; $\bar{d}_i$ is the average Euclidian distance between each point in the ith cluster and the centroid of the ith cluster; $\bar{d}_j$ is the average Euclidian distance between each point in the jth cluster and the centroid of the jth cluster; d_{ij} is the Euclidean distance between the centroids of the ith and jth clusters. The maximum value of D_{ij} represents the worst-case within-to-between cluster ratio for cluster i. The optimal clustering solution has the smallest Davies–Bouldin score value.

The significance of cluster separation was assessed by comparing the *DBS* of the wildtype sequences to the randomized scores obtained from 100 permutations of gene group labels (early/late).

In order to visualize the clustering, a principal component analysis (*PCA*) was applied to project the *RSCF* vectors to a plane spanned by their first two principal components. In order to visualize the separation between clusters a maximum margin separation line—a line for which the distance between it and the nearest point from either of the groups is maximized, was calculated and plotted.

In the same manner, analysis of clustering between early and late viral groups and a set of top 50 host genes with the highest protein abundance can be performed.

2.4.1. Mean typical decoding rate (*MTDR*)

A measure which estimates the translation elongation efficiency of the entire gene as a geometric average of TDRs of its codons:

$$MTDR=e^{\left(1/L\right)\sum_{l=1}^{L}\log\left(TDR_i(l)\right)},$$

where i is an index of a codon and L is the gene length in codon unit.

We computed the *MTDR* based on the ribosome profiling data of the *E. coli* (TDRs are effected by translation factors which are common to the host and the bacteriophage). Thus, it should not be sensitive to the low number of bacteriophage reads at the first time points.

2.4.2. Relative translation elongation efficiency coefficient (RTEC)

Quantifies the relative differences in mean *MTDR* values of early and late gene: $RTEC=\frac{(mean\ \ MTDR_E\ \ -mean\ \ MTDR_L)}{(mean\ \ MTDR_E\ \ +mean\ \ MTDR_L)}$ where E and L sigh for early and late genes.

2.4.3. Folding energy analysis

Minimum free folding energy (*MFE*) is a thermodynamic energy involved in maintaining a secondary structure available to perform physical work whilst being released, and thus is characterized by non-positive values. mRNA secondary structure is believed to be in the most stable conformation when minimum amount of free energy is exerted (the *MFE* obtains the most negative value). The local *MFE*-profiles were constructed by applying a 39 nt length sliding window to a genomic sequence: in each step the *MFE* of a local subsequence enclosed by the corresponding window was calculated by Vienna (v. 2.1.9) package RNAfold[31] function with default parameters (see, also[16,32]). This function predicts the MFE and the associated secondary structure for the input RNA sequence using a dynamic programming based on the thermodynamic nearest-neighbor approach (the Zucker algorithm).[33]

First, all the genes in the bacteriophage genome were lined up according to their start codon and *MFE*-profiles were calculated for each coding region together with 40-nt up-stream the start codon sequence from the 5′ UTR. Then, all the genes in the genome were lined up according to their end codon and *MFE* profiles were calculated for each coding region together with 40-nt down-stream the end codon from the 3′ UTR.

For each gene 100 randomized *MFE*-profiles variants were computed basing on randomized sequences generated by the dinucleotide preserving and codon preserving randomization models (both preserving also the encoded proteins, see section above).[16,17,28] We did not change the UTRs in the randomization as in this study we are interested in the coding regions.

The mean *MFE*-profile was obtained by averaging the *MFE*-profiles of all genes (in a position wise manner). In a similar manner, 100 randomized mean *MFE*-profiles were computed by grouping the randomized *MFE*-profiles of all genes in 100 groups, each group contains a different variant for each gene, and then averaging the profiles in each group in a position-wise manner.

In order to assess the statistical significance of the folding strength at a particular position in a sequence, we compared the mean *MFE* values at this position with the mean *MFE* values in the corresponding position in each one of the randomized variants by calculating an empiric *P*-value—a proportion of the randomized values as extreme as in the wild type. Positions with *MFE* related *P*-value <0.05 were defined as selected for strong/weak folding.

In addition, we computed mean *MFE* values for each gene over all windows (by averaging the values in the corresponding *MFE*-profiles) and compared them to the mean *MFE* values obtained from the corresponding 100 randomized profiles. For each gene we calculated its mean *MFE* value and an average of 100 mean *MFE* values from its randomized variants; the distributions of the wildtype and randomized mean *MFE* values of different genes were compared using Wilcoxon signed-rank test. Early and late genes were analysed separately.

2.4.4. Average Repetitive Substring (*ARS*) index

This measure is based on the assumption that evolution shapes the organismal coding sequences (and other part of the gene) to improve their interaction with the intra-cellular gene expression machinery. Since these interactions are mediated via binding of the gene expression machinery (e.g. translation/transcription factors, RNAP, ribosomes, RNA binding proteins, etc.) to the genetic material (DNA or RNA), the genetic material tend to have optimized binding sites (which are sub-sequences of nucleotides). We also expect that binding sites will appear in many coding regions and that more optimal binding sites will tend to appear more times in the genome. Thus, if longer substrings of a genome tend to appear in a certain organism's coding sequence, it suggests that this coding sequence is more optimized to the intra-cellular gene expression machinery and thus it is probably more highly expressed. Here, we computed the *ARS* index for each bacteriophage gene in comparison to the host (*E. coli*) and in comparison to the rest of the viral genes.

The algorithm of *ARS index* is based on the following steps: (i) For each position i in the coding sequence S find the longest substring S_i^j that starts in that position, and also appears in at least one of the coding sequences of the reference genome (*E. coli*/viral). (ii) Let $|S|$ denote the length of a sequence S; the *ARS index* of S is the mean length of all the substrings S_i^j : $ARS = \frac{\sum_{S_i^j}}{|S|}$.

2.4.5. Rare codons analysis

Rare codons in a reference set of coding sequences were defined as codons with the relative synonymous frequency <0.2. Three reference sets were used: *E. coli* coding sequences, bacteriophage early coding sequences, bacteriophage late coding sequences.

A rare codons score (*RCS*) for a specific early/late/*E. coli* coding sequence with respect to a reference set of all early/all late/all *E. coli* coding sequences is defined as a percentage of amino acids in that sequence encoded by a rare codon out of all amino acids that are encoded by at least one rare codon in the corresponding reference set (if an AA is not encoded by codons that are rare in the reference set we exclude it from the analysis):

$$RCS = \frac{1}{N}\sum_c I_c,\ I_c = \begin{cases} 1, & RSCF(c) < 0.2 \\ 0, & otherwise \end{cases},$$

where the sum is over all codons c that have at least one rare synonymous codon that appears in the reference set and N is the total number of such codons.

2.4.6. Late genes sampling

In order to control the influence of the difference of genes length between early and late groups we sampled the late genes so that the average length of both genes groups is the same. The sampling was of random contiguous blocks of codons from the late genes and according to distributions of early genes.

3. Results and discussion

The research outline of the study is described in Fig. 1A. Our analysis was based on the genome (mainly the coding sequences) of the *E. coli* host (A), the genome of bacteriophage lambda (B) and the ribo-seq measurements of these two (C). To assess the statistical significance of the signals found in the analysed viral genes and to exclude the possibility that these signals are un-direct consequences of other genomic properties, we compared them to signals expected by chance in the corresponding randomized variants (D); two different randomization models were employed: one that maintains the encoded proteins and the frequencies of pairs of adjacent nucleotides (dinucleotides), and the other that maintains the encoded proteins and the frequencies of synonymous codons (codon usage bias). Basing on the ribo-seq data, the expression levels of each gene at each time point (E), the relative decoding rate of each codon (F), and the classification of the bacteriophage genes to early and late (with respect to the beginning of the lytic phase) (G) were derived. On the basis of A, B, D, G, we performed synonymous codons usage analysis of coding regions in viral and *E. coli* genes (H) using the *RSCF* (I) and tRNA adaptation indexes (*tAI*) (J). On the basis of E, F, G, we analysed codons decoding rates for early/late genes at different stages of the viral development (L) on gene/genomic (M) and per-codon (N) levels. In addition, based on A, B, D, G, we studied the local and global signals of evolutionary selection for strong/weak mRNA folding (K) and for higher order synonymous information encoded in repetitive sequence motifs that are longer/more complex than single codons in the coding regions of the bacteriophage genes (O).

The major aim of this study is to compare the properties of coding regions of bacteriophage early and late genes; thus, we start with a brief description of the expression pattern of these two gene groups. The analysis of ribo-seq read count per nucleotide for early and late groups of bacteriophage genes appears in Fig. 1B. As can be seen, at the first time point the read count of both groups is very low.

Afterwards the expression levels of both gene groups increase; whilst the expression of the early gene group is dominant during minutes 1–10, the expression levels of the late group become dominant towards the 20th minute. Specifically, 32 of the bacteriophage genes are defined as early genes as their expression levels increase from 5 to 10 min, and then decrease by minute 20 after the beginning of the lytic process[9]; the rest of the genes (34 genes) are define as late genes as their expression levels become significant only at minute 10 from the beginning of the lytic process, and then increase considerably by minute 20 (Fig. 1B and[9]).

3.1. Evidence of selection for different codons in early and late genes

At the first step (Fig. 2A), we aimed at comparing the synonymous codons usage in *E. coli* and bacteriophage early and late coding sequences. To this end, for each coding sequence, we computed its relative synonymous codon frequencies (*RSCF*)—a 61-dimensional vector representing each codon (except the stop codons) by its frequency in that sequence normalized relative to the frequencies of other synonymous codons coding for the same amino acid (Methods). Our analysis demonstrated that the early and late genes tend to be clustered into two significantly separated (P-value <0.01) clusters according to their synonymous codons usage. In addition synonymous codon usage in both groups of viral genes was found to be significantly different (P-value <0.01) from that of *E. coli* (Fig. 2A; Methods). These results provide evidence that different sets of synonymous codons for early vs. late genes are selected in the course of viral evolution and may be related to the optimization of the viral fitness.

3.2. Differential codon usage in early and late genes can be partially explained by adaptation of translation elongation efficiency to different bacteriophage development stages

Having shown that early and late viral genes have a significantly different composition of synonymous codons which may be associated with various features of their expression, we would like to focus one such feature, and understand the translation elongation efficiency of bacteriophage coding regions and how it behaves in different stages of the viral lytic cycle.

To this aim we employed a *condition* specific measure of translation elongation[30] to study the elongation speed of viral codons/genes during the different steps of phage development. This measure, called *MTDR*, is based on the estimation of a typical codon decoding rate (*TDR*) of each codon at each time point[30] based on the ribosome profiling data[9] and enables ranking codons and coding regions according to their elongation rate whilst controlling for other factors, such as initiation rates and mRNA levels (see details in the Methods section).

At the first step, we wanted to check whether the bacteriophage coding regions undergo any selection for optimizing translation elongation. To this end, at every time condition we computed two average *MTDR* values, for early and late genes separately, and compared them to the average *MTDR* values obtained for corresponding randomized variants that maintain the wild type amino acid content and the dinucleotide distribution (Methods). As can be seen in Fig. 2B and Supplementary Fig. S1, the average *MTDR* of both groups is significantly higher than expected from the random model in all time points (early: $p<0.01$; late: $p<=0.04$). These results suggest that, indeed, translation elongation efficiency is maintained along the lytic cycle of infection and may be a factor that drives codon evolution in both groups of the bacteriophage genes.

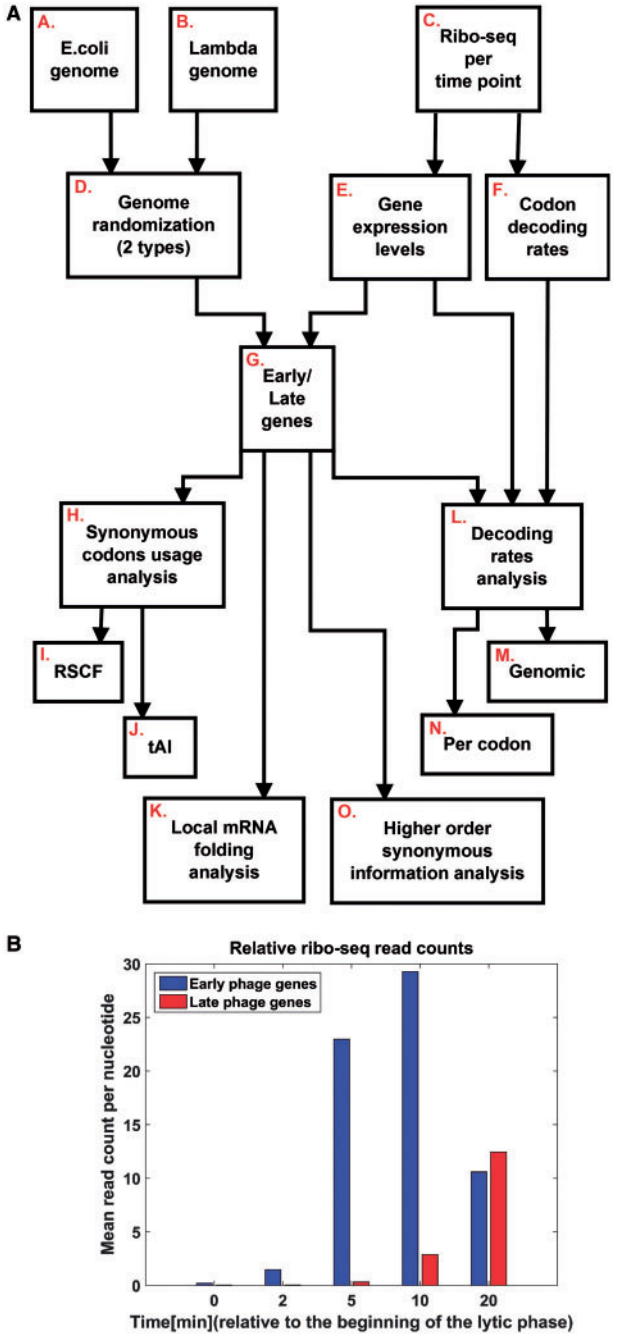

Figure 1. (A) A flow diagram and illustration of the study (see details in the main text). (B) Relative expression level of each of the gene groups (early/late) in read count per nucleotide.

At the next step, to compare the translation elongation efficiency between early and late genes, we looked at the *RTEC* which quantifies the relative differences in average *MTDR* values of two gene groups: more positive *RTEC* values mean that early genes are more efficient than the late genes and vice versa, more negative *RTEC* values mean that late genes are more efficient than the early ones; *RTEC* values close to zero mean that the two groups of genes have a

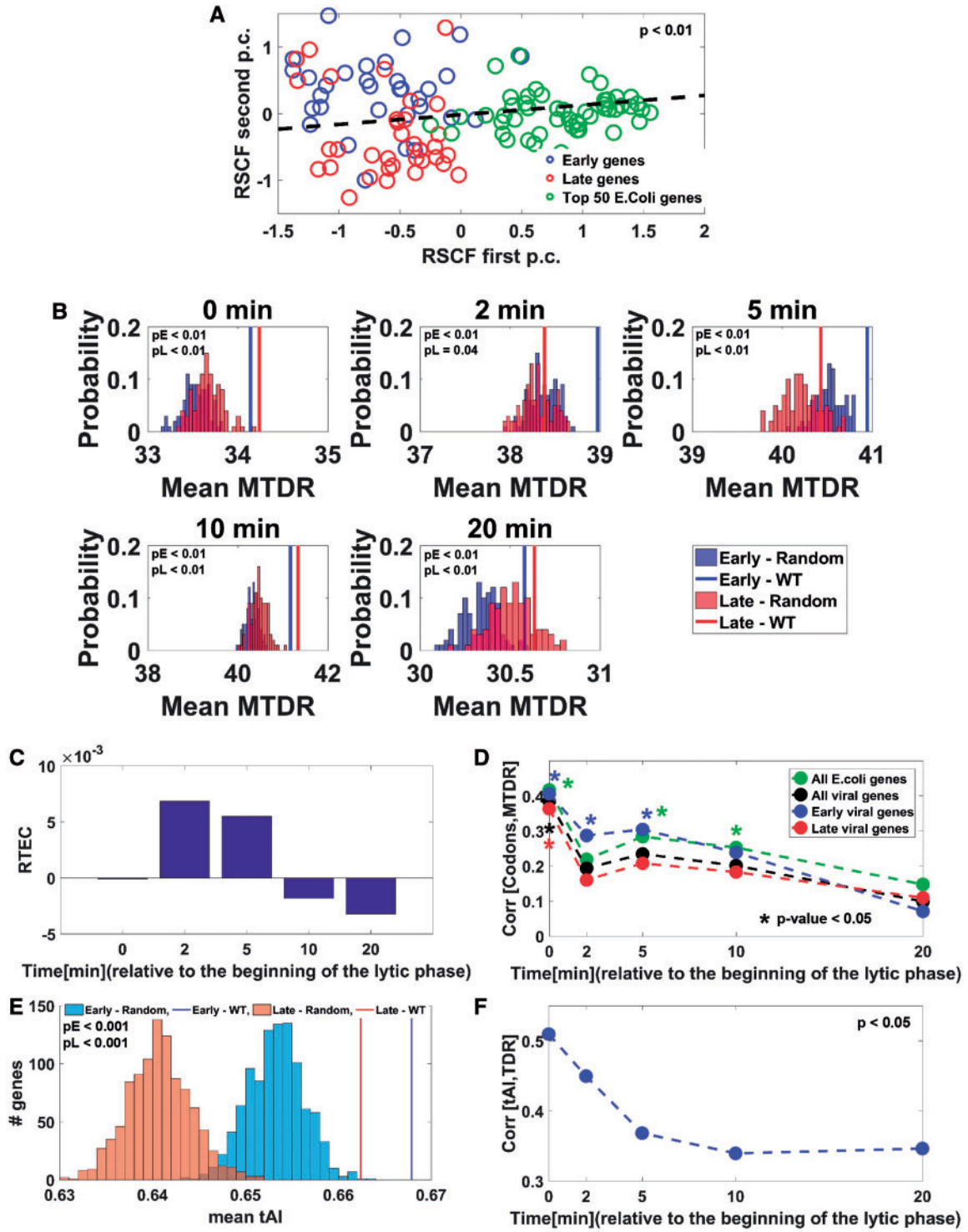

Figure 2. (A) Selection for different codons in early and late genes. A plot of two principal components of RSCF vectors of all bacteriophage and *E. coli* genes. Bicteriophage early (blue) late (red) and *E. coli* (green) genes tend to be clustered into two distinct groups according to their synonymous codons compositions. The separation between the groups was measured by Davies–Bouldin index and found to be significantly higher than expected in random ($P < 0.01$; see methods). The separation between the groups of early and late viral genes was visualized by a maximum margin separation line—a line for which the distance between it and the nearest point from either of the groups is maximized. (B) Selection for translation elongations efficiency in bacteriophage coding regions. At each time point: average MTDR values of wildtype early/late genes (vertical bars) were compared with MTDR values of 100 corresponding randomized variants (histograms). Average wildtype MTDR values of each group are significantly higher ($P < 0.05$) than expected in random. The late genes were sampled to control the length factor (see methods; see also Supplementary Fig. S1). (C) Adaptation of translation elongation efficiency in early and late genes to different bacteriophage development stages genes. Relative translation elongation efficiency coefficient, $\mathrm{RTEC} = (mean\ MTDR_E - mean\ MTDR_L)/(mean\ MTDR_E + mean\ MTDR_L)$, as a function of time from the beginning of the lytic stage (0–20 min), where $MTDR_E$ and $MTDR_L$ sigh for the MTDR of early and late genes, respectively. We can see that the RTEC of early genes is higher at the beginning and become lower with time (as expected); the first point ($t = 0$), when there are no measurements of expression, is ignored (see also Fig. 1B). (D) Correlation between

similar translational efficiency (Methods). Figure 2C describes the *RTEC* as a function of time (0–20 min). As can be seen, the relative efficiency of elongation of early genes (in comparison to the late genes) is high at the beginning and become lower with time (*P*-value = 0.04; based on spearman correlation). These results demonstrate that translation elongation efficiency of the early genes is relatively higher at the early stages of the bacteriophage development (when they are expressed) and the translation elongation efficiency of the late genes is relatively higher at the late stages of the bacteriophage development (when they are expressed).

Figure 2D describes the per codon correlation between the *TDR* and *RSCF* for the two bacteriophage gene groups (early and late) and the *E. coli* genes at different time points. As can be seen, the correlation is higher and significant for the first time points in the case of the *E. coli* and early bacteriophage genes. For the late genes, the correlation is significant only at the initial points. The fact that the correlation between *RSCF* and *TDR* in the case of the *early viral genes is* significant at the initial points supports the conjecture that the relative codon decoding times change during the viral development; this is probably related, among others, to the fact that the bacteriophage affects extensively the gene expression in the cell.

The lower correlations at time 2 does not seems to be relate to trivial biases/problems with the experiment as the number of reads in the *E. coli* (used for inferring the *MTDR*) is similar to the number of read in the different time points; in addition, the number of reads mapped to the viral genes is higher than in time 0. Thus, it is possible that the lower correlation is related to a biological phenomenon: e.g. it is possible that in this time point there is (strong) deviation (which is possibly short in time) of the concentration levels of the translation factors in the cell related to the other points; whilst the codon distributions were shaped to fit the other (longer) periods of the bacteriophage development.

3.3. Selection for translation efficiency in bacteriophage genes may be partially explained by adaptation to the *E. coli* tRNA pool, and the fact that it changes during the bacteriophage development

Previous studies demonstrated that that the codon decoding times may be directly influenced by the tRNA levels in the cell (see, for example[30]).

In Fig. 2E, we analysed the adaptation of the viral codons to the genomic tRNA copy number in the host at natural conditions (Methods) and found it to be significant in comparison to the randomized variants that maintain the dinucleotides distribution for both early and late gene groups. However, as can be seen in Fig. 2F, the correlation between the *TDR* and the *tAI* of different codons is significant but *decreases* during the viral development stages. These results may suggest that among others the tRNA levels change during the viral development stages, affecting the codon decoding times.

In addition, we analysed the tendency of early/late bacterial coding sequences to use rare synonymous codons with respect to early/late/bacterial gene groups, respectively (RCS). Supplementary Fig. S2 describes the per condition partial correlation (controlling for gene length) between RCS and mean read counts for the two viral gene groups (early and late) and the *E. coli* genes. As can be seen, the correlation decreases in the case of the early genes and *E. coli* genes and increases for late genes. Our analysis demonstrates that early/late genes with rare early/late genes codons tend to be lowly expressed at the early/late stages, respectively.

The results reported in this section support the conjecture that some of the differences between the early and the late genes are related to the adaptation of viral codons to the intracellular environments in different stages of the phage development. Specifically, we suggest that such adaptation may be result of the fact that the typical decoding times (possibly due to changes in tRNA levels) change during the bacteriophage development.

3.4. Additional constraints that shape the codon content of the bacteriophage

In the previous section, we emphasized the importance of the translation elongation efficiency on shaping the codon composition of the Bacteriophage coding regions. In this section, we demonstrate that additional gene expression aspects are also encoded (in parallel) in the coding sequences.

First, we examined whether there is a selection for strong/weak local folding in different regions along the genomes (coding regions and flanking UTRs) (Fig. 3A–C). To this end, all early/late genes were aligned to the start codon and average values of a minimal folding energy (*MFE*) over all genes in each group were predicted within all possible 39-nt length local window. These average *MFE* values where compared in a position wise manner (Fig. 3A and B) and in average over all positions (Fig. 3C and Supplementary Fig. S3) to the *MFE* values expected in random. Positions with significantly strong (more negative *MFE*) or weak (more positive *MFE*) folding where identified (Fig. 3A and B) and also global *P*-value related to the mean average *MFE* values was computed (Fig. 3C). To make sure that the obtained folding signals were not mainly an indirect consequence of codon usage bias and/or selection for specific dinucleotide contents not related to mRNA folding we employed two randomization models, one designed to maintain both the encoded protein and the distribution of dinucleotides and the other designed to maintain both the encoded protein and the codon usage bias (see the Methods section for more details).

Our analysis suggests that in general, there is an evidence of evolutionary selection for strong folding in the coding region of the bacteriophage in the case of the late genes but not in the case of the early genes and this is not related to very specific region along the coding region (Fig. 3A–C).This selection is manifested by lower average *MFE* values than expected in random (Fig. 3C and Supplementary Fig. S3) and a higher number of positions selected for strong folding

codons typical decoding rates (TDR) and relative synonymous codons frequencies (RSCF) at different time conditions for all, early and late viral genes and all *E. coli* genes. Time points with significant correlations (Spearman *P*-values <0.05) are marked by asterix. For early genes, the correlation is higher than for late and *E. coli* genes and is significant (*P*-value <0.05) for the firsttime points. No significant correlation can be seen for late genes except the first time point. The correlation in the case of the *E. coli* is significant up to 10 min (except at 2 min). (E) Selection for adaptation to *E. coli* tRNA pool in both early and late genes. Average tAI values of wildtype early (blue)/late (red) genes (vertical bars) were compared with tAI values of 1,000 corresponding randomized variants (histograms). Average wildtype tAI values of each group are significantly higher ($P< 0.001$) than expected in random. The late genes were sampled to control the length factor (see Methods). (F) Correlation between TDR and tAI values for each codons at different type points is significant (*P* value < 0.05).

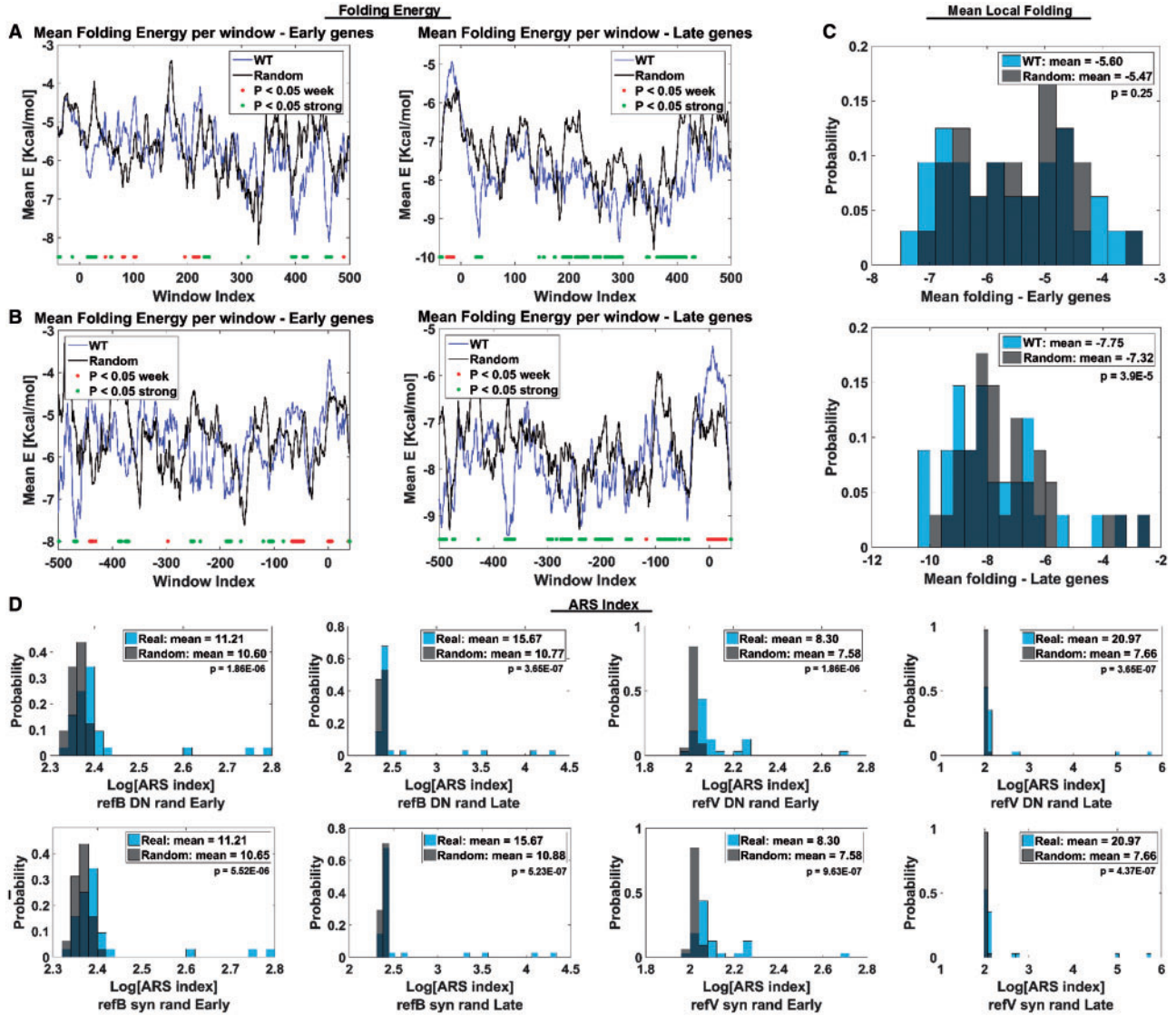

Figure 3. (A and B) Profiles of folding energy (average MFE in all windows of 39-nt length) across the bacteriophage genome (blue) vs. an averaged profile corresponding to 100 randomized variants (black) based on dinucleotide preserving randomization; the window index denotes the distance (in nucleotides) from the beginning of the ORF to the beginning of the window. Regions where the folding energy of the wild type genome is significantly higher (red) or lower (green) than in randomized variants are marked at the bottom of the figure. (A) The profiles include the 5′-UTR near the beginning of the ORF (negative window indexes). (B) The profiles include the 3′-UTR near the ending of the ORF (positive window indexes). (C) Histograms of mean local folding energies (folding energies averaged over all the windows of each gene) compared with randomized mean local folding energies obtained from two models: (i) protein + dinucleotides preserving and (ii) protein + codon usage bias preserving (see also Supplementary Fig. S3). (D) Histograms of log[ARS index]. Eight analyses were performed: two types of reference genomes; bacterial and viral, two type of randomizations; dinucleotide and codons, two groups of genes; early and late. In each histogram, the wild type distribution is compared with the mean random distribution (1,000 random genomes). The *P*-values were calculated according to Wilcoxon signed-rank test.

(27 in the early and 123 in the late genes in Fig. 3A; 37 in the early and 144 in the late genes in Fig. 3B).

Local genomic regions (both in UTRs and in coding sequences) selected for strong mRNA folding may be related to various functional mRNA structures in the coding regions that effect the viral life cycle[16,17,34]; specifically, it is possible that the mRNA structures are more important in the late genes due to less canonical regulatory aspects needed at the later phases of viral life. Interestingly, we found that in the case of the late genes there is an evidence of selection for weak folding at the 5′ end of the coding regions, suggesting that there is a significant evolutionary pressure for improved initiation efficiency.[34]

Secondly, we performed an analysis with a measure that detects long sub-sequences in the viral genome (*ARS index*) and were able to detect gene expression codes in an unsupervised manner[35] (see technical details in the Methods section). Our analysis (Fig. 3D) demonstrates that both in the early and late genes there are significant gene expression codes in the bacteriophage coding regions that cannot be explained by the viral codon frequency or dinucleotide composition. These results suggest that in parallel to single codon adaptations, the bacteriophage coding regions undergoes an evolutionary pressure to include more complex/longer regulatory signals in the coding region. This suggests that additional future studies should be performed on this topic.

4. Conclusions

The analyses performed in this study emphasize the way evolution shape the coding region of Bacteriophage Lambda to improve its expression levels and fitness. We demonstrate that evolution shapes the codon content of the Bacteriophage genes to fit them to the dynamic intracellular environment during the Bacteriophage development. Specifically, we show that the codon frequencies of the early and late genes were shaped in a different manner: (i) there is high intra group similarity in codon usage frequencies than inter-group similarities; (ii) the codons frequencies of both groups are significantly different than the codon frequencies of the host; (iii) whilst the decoding rates of both group is higher than expected based on the amino acid content and dinucleotide composition, the mean decoding rate of the early genes is relatively higher/lower than the late genes during the initial/late stages of the viral development, respectively.

The results reported here demonstrate and suggest that it may be possible to better understand the function of viral genes via the analysis of their codon distribution. They can also promote developing novel approaches for vaccine development and viral therapies: For example, based on the codon frequencies of different viral ORF we may be able to predict if they are early of late genes. In addition, engineering the viral genome such that its decoding rate (e.g. as measured by ribosome profiling) is attenuated may enable generating live attenuated based vaccines; on the other hand, improving the viral decoding rate may contribute to developing efficient oncolytic viruses and contribute towards developing efficient procedures for generating inactivated vaccines.

Our analysis support the conjecture that the that tRNA pool in the host changes during the bactriophage development with a tRNA pool similar to the host pool in natural conditions at the early points but more different than the host natural tRNA pool during the late points (see Fig. 2F). Thus, the accurate measure for estimating the optimization of the codon decoding rates of the bactriophage genes is the *TDR* and measures based on the genomic tRNA copy number (such as the *tAI*).

Nevertheless, it is important to mention that previous studies suggest that relative levels of tRNAs tend to have high correlation among different conditions and tissues (see for example[36,37]). This may be related, among others, to the fact that a significant change of tRNA levels ranking may have strong effect on co-translation protein folding (see for example[38]) and thus a strong effect on the functionality of many proteins, effecting both the host and the bacteriophage. Similar phenomena was observed also in this study—there is significant correlation between codon decoding rates at different time points and the relevant genomic tRNA copy number (see Fig. 2F); however, as mention, this correlation is lower at the later time points.

At each time point the 'optimal' solution relate to the optimization of translation elongation speed and the optimal allocation of translation factors include using the codons with the highest codon decoding rates (i.e. the ones related to factors such as tRNA molecules that have relatively higher abundance in the cell). Thus, genes undergo selection to have codons similar to the translation factor concentration and this selection is expected to be higher in highly expressed genes (see[39,40]) in each condition since non-optimality in these genes should have higher effect (in comparison to lowly expressed genes) on the organism/viral fitness; eventually, at early/late stages the early/late genes are expected to have higher mean decoding rate than late/early genes, respectively (as we see in Fig. 2C).

Whilst the codons of the early/late genes are different both groups use all the codons. However, the fact that the two groups of genes have different codons also means that they use a little different set of tRNA molecules—the late genes use tRNA molecules less used by the early genes.

An interesting topic for future study is measuring the tRNA levels in different time point and an interesting topic for future study will be to perform such measurements to validate and better understand the different codon frequencies of early/late genes.

Finally, whilst this study focused on codon frequencies and its relation to translation elongation our analysis suggest that additional regulatory aspects are encoded in via the local folding of the viral RNA and possibly additional viral genomic motives (longer than singe codon; Fig. 3D); these results and the previous ones demonstrate that the complexity and information content of the Bacteriophage genome is higher than thought before and encourages further studies in this direction.

Supplementary data

Supplementary data are available at *DNARES* Online.

Acknowledgements

E.G. and A.D. are supported, in part, by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv Universit. A.D. is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship. T.T. is grateful to the Minerva ARCHES award.

Conflict of interest

None declared.

References

1. Haq, I.U., Chaudhry, W.N., Akhtar, M.N., Andleeb, S. and Qadri, I. 2012, Bacteriophages and their implications on future biotechnology: a review. *Virol. J.*, **9**.
2. Clark, J.R. and March, J.B. 2006, Bacteriophages and biotechnology: vaccines, gene therapy and antibacterials. *Trends Biotechnol.*, **24**, 212–218.
3. Hermoso, J.A., Garcia, J.L. and Garcia, P. 2007, Taking aim on bacterial pathogens: from phage therapy to enzybiotics. *Curr. Opin. Microbiol.*, **10**, 461–472.
4. Bull, J.J., Cunningham, C.W., Molineux, I.J., Badgett, M.R. and Hillis, D.M. 1993, Experimental molecular evolution of bacteriophage-T7. *Evolution*, **47**, 993–1007.
5. Arias, C., Weisburd, B., Stern-Ginossar, N., et al. 2014, KSHV 2.0: a comprehensive annotation of the Kaposi's sarcoma-associated herpesvirus genome using nextgeneration sequencing reveals novel genomic and functional features. *Plos Pathog.*, **10**.
6. Yang, Z.L., Cao, S., Martens, C.A., et al. 2015, Deciphering poxvirus gene expression by RNA sequencing and ribosome profiling. *J. Virol.*, **89**, 6874–6886.
7. Stern-Ginossar, N. and Ingolia, N.T. 2015, Ribosome profiling as a tool to decipher viral complexity. In: L.W., Enquist, (ed.), *Annual Review of Virology, Vol. 2*, pp. 335–349.
8. Irigoyen, N., Firth, A.E., Jones, J.D., Chung, B.Y.W., Siddell, S.G. and Brierley, I. 2016, High-resolution analysis of coronavirus gene expression by RNA sequencing and ribosome profiling. *Plos Pathog.*, **12**.
9. Liu, X.Q., Jiang, H.F., Gu, Z.L. and Roberts, J.W. 2013, High-resolution view of bacteriophage lambda gene expression by ribosome profiling. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 11928–11933.
10. Carbone, A. 2008, Codon bias is a major factor explaining phage evolution in translationally biased hosts. *J. Mol. Evol.*, **66**, 210–223.

11. Coleman, J.R., Papamichail, D., Skiena, S., Futcher, B., Wimmer, E. and Mueller, S. 2008, Virus attenuation by genome-scale changes in codon pair bias. *Science*, **320**, 1784–1787.
12. Jenkins, G.M., Pagel, M., Gould, E.A., Zanotto, P.M.D. and Holmes, E.C. 2001, Evolution of base composition and codon usage bias in the genus Flavivirus. *J. Mol. Evol.*, **52**, 383–390.
13. Shackelton, L.A., Parrish, C.R. and Holmes, E.C. 2006, Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. *J. Mol. Evol.*, **62**, 551–563.
14. Herskowi I. 1973, Control of gene-expression in bacteriophage-lambda. *Ann. Rev. Genet.*, **7**, 289–324.
15. Casjens, S.R. and Hendrix, R.W. 2015, Bacteriophage lambda: early pioneer and still relevant. *Virology*, **479**, 310–330.
16. Goz E.T.T. 2016, Evidence of a direct evolutionary selection for strong folding and mutational robustness within HIV coding regions. *J. Comput. Biol.*, **23**.
17. Goz, E. and Tuller, T. 2015, Widespread signatures of local mRNA folding structure selection in four Dengue virus serotypes. *BMC Genom.*, **16**.
18. Wang, Q., Barr, I., Guo, F. and Lee, C. 2008, Evidence of a novel RNA secondary structure in the coding region of HIV-1 pol gene. *RNA-Publ. RNA Soc.*, **14**, 2478–2488.
19. Alvarez, D.E., Ezcurra, A.L.D., Fucito, S. and Gamarnik, A.V. 2005, Role of RNA structures present at the 3′ UTR of dengue virus on translation, RNA synthesis, and viral replication. *Virology*, **339**, 200–212.
20. Smith, D.B., Mellor, J., Jarvis, L.M., et al. 1995, variation of the hepatitis-C virus 5′ noncoding region – implications for secondary structure, virus detection and typing. *J. Gener. Virol.*, **76**, 1749–1761.
21. Jenkins, G.M. and Holmes, E.C. 2003, The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Res.*, **92**, 1–7.
22. Sharp, P.M., Rogers, M.S. and McConnell, D.J. 1985, selection pressures on codon usage in the complete genome of bacteriophage-T7. *J. Mol. Evol.*, **21**, 150–160.
23. Sahu, K., Gupta, S.K., Sau, S. and Ghosh, T.C. 2005, Comparative analysis of the base composition and codon usages in fourteen mycobacteriophage genomes. *J. Biomol. Struct. Dyn.*, **23**, 63–71.
24. Sau, K., Gupta, S.K., Sau, S. and Ghosh, T.C. 2005, Synonymous codon usage bias in 16 *Staphylococcus aureus* phages: implication in phage therapy. *Virus Res.*, **113**, 123–131.
25. Kunisawa, T., Kanaya, S. and Kutter, E. 1998, Comparison of synonymous codon distribution patterns of bacteriophage and host genomes. *DNA Res.*, **5**, 319–326.
26. MARTIN, M. 2011, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.*, **17**, 2226–6089.
27. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. 2009, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**.
28. Zhang, Y., Ponty, Y., Blanchette, M., Lecuyer, E. and Waldispuhl, J. 2013, SPARCS: a web server to analyze (un)structured regions in coding RNA sequences. *Nucl. Acids Res.*, **41**, W480–W485.
29. dos Reis, M., Savva, R. and Wernisch, L. 2004, Solving the riddle of codon usage preferences: a test for translational selection. *Nucl. Acids Res.*, **32**, 5036–5044.
30. Dana, A. and Tuller, T. 2014, The effect of tRNA levels on decoding times of mRNA codons. *Nucl. Acids Res.*, **42**, 9171–9181.
31. Lorenz, R., Bernhart, S.H., Siederdissen, C.H.Z., et al. 2011, ViennaRNA package 2.0. *Algorithms Mol. Biol.*, **6**.
32. Tuller, T., Waldman, Y.Y., Kupiec, M. and Ruppin, E. 2010, Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 3645–3650.
33. Zuker, M., and Stiegler, P 1981, Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.*, **9**, 133–148.
34. Tuller, T. and Zur, H. 2015, Multiple roles of the coding sequence 5′ end in gene expression regulation. *Nucl. Acids Res.*, **43**, 13–28.
35. Zur, H. and Tuller, T. 2015, Exploiting hidden information interleaved in the redundancy of the genetic code without prior knowledge. *Bioinformatics*, **31**, 1161–1168.
36. Tuller, T., Carmi, A., Vestsigian, K., et al. 2010, An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, **141**, 344–354.
37. Mahlab, S., Tuller, T. and Linial, M. 2012, Conservation of the relative tRNA composition in healthy and cancerous tissues. *RNA-Publ. RNA Soc.*, **18**, 640–652.
38. Yu, C.H., Dang, Y.K., Zhou, Z.P., et al. 2015, Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. *Mol. Cell*, **59**, 744–754.
39. Sabi, R. and Tuller, T. 2014, Modelling the efficiency of codon–tRNA interactions based on codon usage bias. *DNA Res.*, **21**, 511–525.
40. dos Reis, M. and Wernisch, L. 2009, Estimating translational selection in eukaryotic genomes. *Mol. Biol. Evol.*, **26**, 451–461.