Virulent but not temperate bacteriophages display hallmarks of rapid translation initiation

Adam J. Hockenberry[1,*], David C. Weaver[1], Claus O. Wilke[1]

1 Department of Integrative Biology, The University of Texas at Austin

* Corresponding author

Bacteriophages rely almost exclusively on host-cell machinery to produce their proteins, and their mRNAs must therefore compete with host mRNAs for valuable translational resources. In many bacterial species, highly translated mRNAs are characterized by the presence of a Shine-Dalgarno sequence motif upstream of the start codon and weak secondary structure within the start codon region. However, the general constraints and principles underlying the translation of phage mRNAs are largely unknown. Here, we show that phage mRNAs are highly enriched in strong Shine-Dalgarno sequences and have comparatively weaker secondary structures in the start codon region than host-cell mRNAs. Phage mRNAs appear statistically similar to the most highly expressed host genes in *E. coli* according to both features, strongly suggesting that they initiate translation at particularly high rates. Interestingly, we find that these observations are driven largely by virulent phages and that temperate phages encode mRNAs with similar start codon features to their host genes. These findings apply broadly across a wide-diversity of host-species and phage genomes. Further study of phage translational regulation—with a particular emphasis on virulent phages—may provide new strategies for engineering phage genomes and recombinant expression systems more generally.

Introduction

Protein translation consumes a large amount of cellular resources, and the genomes of microbial species show strong evidence of selection for rapid and efficient translation [1–8]. The statistical analysis of genome sequence features has provided important insights into many of the basic mechanisms of gene expression, particularly by contrasting sequence patterns within genes that have high- and low-expression demands [9–14] . Over-representation of a purine-rich motif upstream of bacterial start codons, for instance, ultimately lead to the discovery of the Shine-Dalgarno sequence mechanism—a direct binding interaction between the 30S ribosomal sub-unit and mRNA that governs start codon recognition and facilitates translation initiation [15–21]. Collectively, gene sequence analyses have facilitated advances in recombinant protein engineering by deciphering the sequence rules for optimal expression, which have subsequently been validated, refined, and expanded upon in numerous experimental and evolutionary studies [22–32]. However, the amount of available information contained within individual bacterial genomes is limited both in terms of the overall number of proteins and the sequence-level diversity that is present across closely related strains.

Bacteriophage genomes, by contrast, contain a large pool of untapped sequence diversity that metagenomic techniques have only recently begun to characterize in depth [33–39]. Phage mRNAs must be expressed in host cells—in most cases, entirely by existing host cell machinery—and the statistical patterns and constraints that are encoded in these sequences may thus help to further elucidate host-cell transcriptional and translational constraints and mechanisms [40, 41]. While several model phage species are well-characterized [42–46], there have been comparatively few investigations into larger-scale statistical patterns present within phage genomes [47–53]. Deciphering the coding sequence rules governing phage genome evolution is additionally critical for engineering phages as well as for determining the utility of this knowledge for host-cell engineering applications [54, 55].

Here, we analyze thousands of complete, high-quality phage genomes from 33 different bacterial hosts to determine whether the constraints shaping translation-initiation regions in phage mRNAs differ from host-cell genomes. We first present an in depth analysis of *E. coli*-infecting phages and find that phage genomes are predicted to have translation-initiation rates that are on-par with host-cell genes encoding only the mostly highly abundant proteins. Next, we show how translation initiation-related features co-vary and associate strongly with phage lifestyle. Finally, we extend our findings across a broad phylogenetic range of host species, suggesting that phage mRNAs—and virulent phages in particular—are subject to strong evolutionary pressure to ensure rapid translation initiation.

Results

Phage mRNAs are predicted to have rapid translation-initiation rates.

We focused our study on two sequence features that are robustly linked to translation-initiation rates across a wide-range of bacterial species. For a given gene, we first measured the strength of anti-Shine-Dalgarno (aSD) sequence binding interaction (5′-CCUCCU-3′) by selecting the strongest binding hexamer sequence within a narrow window upstream of the annotated start codon (Fig. 1A). Stronger aSD sequence binding (more negative ΔG) is associated with start codon recognition and rapid translation initiation [56]. Next, we calculated the structural accessibility of the start codon by predicting the stability of the secondary structure for a 90 nucleotide fragment surrounding each start codon (30 bases upstream and 60 bases downstream, Fig. 1B). Weaker secondary structure within this region (more positive ΔG) is robustly associated with higher translation-initiation rates [57].

We measured aSD sequence binding and secondary structure strengths for all protein coding genes in phage T7—a well-studied, model phage–and its bacterial host, *E. coli*. We observed a clear distinction whereby the aSD sequence binding strengths of T7 genes are narrower and highly skewed towards stronger binding relative to *E. coli* genes (Fig. 1C, mean of -7.61 vs -4.77 kcal/mol, Welch's t-test $p < 0.001$). The distribution of secondary structure strengths within the start codon region are slightly (but insignificantly) shifted towards weaker secondary structure, which is also associated with higher translation-initiation rates (Fig. 1D, mean of -18.3 vs -19.5 kcal/mol, $p = 0.07$). Given that aSD sequence binding strengths are substantially stronger and that the strength of mRNA secondary structures are statistically similar in phage T7 relative to *E. coli*, this analysis shows that overall translation-initiation rates (TIR) are predicted to be higher for phage T7 mRNAs relative to host mRNAs.

To determine the generality of these findings, we extended this analysis to a set of 254 diverse *E. coli*-infecting phage genomes (see Materials and Methods). Similar to phage T7, we found that 202 of these 254 phage genomes had a stronger mean aSD sequence binding strength when compared to the *E. coli* genome and 82 of these comparisons were significant (Fig. 1E, Welch's t-test with FDR-correction, $p < 0.01$). By contrast, only 10 of the phage genomes displayed significantly weaker aSD binding strengths when compared with *E. coli* (according to the same significance criteria).

When we considered mRNA secondary structure surrounding the start codon, we found that 152 out of 254 tested phage genomes had *weaker* mRNA secondary structure in the start codon region compared with *E. coli* genes and 99 of these comparisons were statistically significant (Welch's t-test with FDR-correction, $p < 0.01$, Fig. 1F). As with aSD sequence binding, this result indicates that phage mRNAs are likely to initiate translation more rapidly than host-cell mRNAs. Only 13 phage genomes had significantly stronger predicted mRNA secondary structures within the start codon region compared to host genome mRNAs (and thus lower predicted translation-initiation rates).

One possible explanation for why phage mRNAs appear to have higher translation-initiation rates than hosts is that phage genomes are highly compact and may be devoid of genes that are rarely expressed or under weak selective pressures—as may be the case for much of the $> 4,000$ protein coding genes within the *E. coli* genome. We therefore repeated the above analyses, comparing phage genomes to subsets of the *E. coli* genome with progressively more stringent cutoffs in terms of their average protein abundances (see Materials and Methods). As expected, we found that *E. coli* mRNAs that encode highly abundant proteins have both substantially stronger aSD sequence binding strengths and weaker mRNA secondary structure in the start codon regions. Achieving parity in terms of roughly balancing the number of phage genomes with significantly stronger (and weaker) aSD sequence binding strengths required considering only the top 10–25% *E. coli* genes with the highest protein abundances (Supplementary Fig. S1A). The results when assessing mRNA secondary structure in the start codon region are similar: overall the translation-initiation region of phage mRNAs appear to be under strong selection that is on par with only the most highly expressed host genes (Supplementary Fig. S1B). We additionally considered essential and non-essential host gene categories separately under the expectation that phage genes might more closely resemble essential host-cell genes. However, we found that phage translation-initiation regions are significantly distinct from both essential and non-essential gene subsets (Supplementary Fig. S1A,B).

Sequence features in translation-initiation regions differ between virulent and temperate phages.

Our findings thus far show that phage mRNAs are predicted to bind strongly to ribosomes and thus are predicted to have increased translation-initiation rates relative to host mRNAs. Additionally, phage mRNAs also appear to have generally weaker secondary structures surrounding the start codon, a feature that is itself associated with rapid recruitment of ribo-

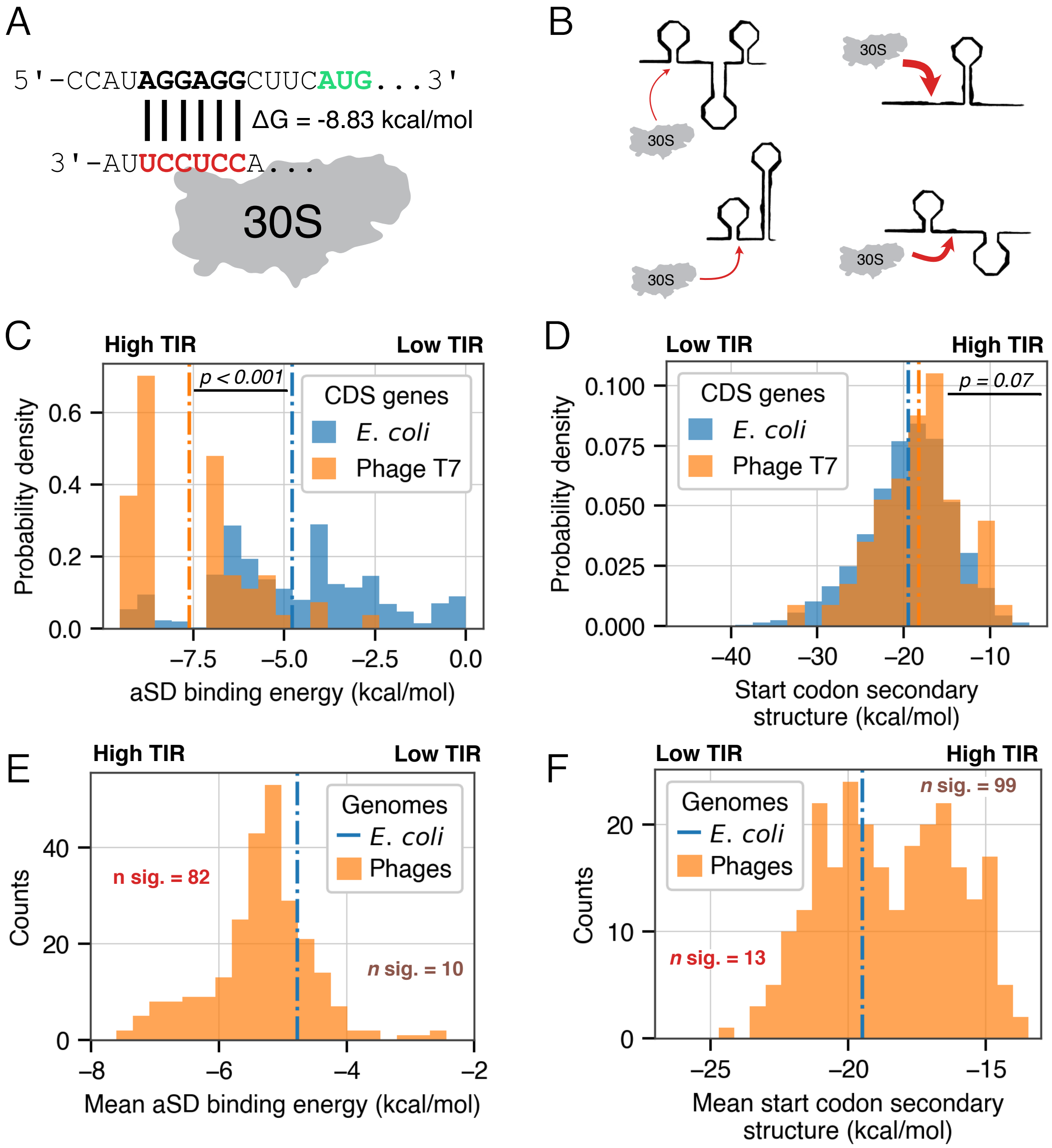


Figure 1. Translation-initiation metrics in phage and host genomes. (A) Illustration of the aSD binding strength, where stronger ribosomal binding is associated with a higher predicted translation-initiation rate (TIR). (B) Illustration of mRNA secondary structure around the start codon, where weak structure is associated with higher predicted TIRs. (C) Histogram of aSD sequence binding strengths upstream of phage T7 and *E. coli* start codons (dashed lines indicate group means). (D) As in (C), showing mRNA secondary structure strength surrounding the start codon. (E) Histogram of mean aSD sequence binding energy across 254 phage genomes. 'n. sig' denotes the number of phage genomes with significant differences compared to *E. coli* (dashed line). (F) As in (E), comparing mean mRNA secondary structure around the start codon.

somes and high translation-initiation rates. However, we have investigated these features in isolation and it is unclear whether these two sequence features represent competing strategies that different phages use to ensure rapid translation initiation or whether these features co-occur within the same genomes.

To investigate the possible differences between phage and host mRNAs along several dimensions, we used a logistic regression framework that attempts to predict the genome of origin (host or phage) based on knowledge of both the aSD sequence binding strength and the strength of mRNA secondary structure surrounding the start codon for each phage-host pairing. Using only a single predictor variable, this procedure is equivalent to a Student's t-test but the flexible nature of the model allows us to analyze both variables simultaneously and to further account for potentially confounding variables. The reported effect size is simply the model coefficient for each variable (a standardized conditional log-odds ratio), the magnitude of which can be directly compared for various predictor variables to assess their overall contribution. We decided to control for two potentially confounding variables: coding sequence GC content and codon usage biases. The GC content of coding sequences can directly influence the strength of mRNA secondary structure (and this feature may vary between phage and host genomes). Additionally, codon usage biases are a general indicator of translational selection—particularly on translation elongation. We found that GC content and codon usage biases do indeed vary between host and phages when these variables are considered in isolation (Supplementary Fig. S2). Our logistic regression model thus contains two features of direct interest related to translation initiation and two features that we treat as potentially confounding variables.

Using the same dataset of 254 *E. coli* infecting phages, we found that strong aSD sequence binding strength and weak mRNA secondary structure co-occur in a large subset of phages (upper-left quadrant, Fig. 2A). The number of significant phages indicated in the figure (37 in the upper-left quadrant) refers to to phages where *both* variables are significantly different from the host (after separately applying a FDR-correction to each variable, $p < 0.01$). However, numerous phage genomes reside in the lower-left quadrant where they display slightly stronger aSD binding strengths than the host genome but also slightly stronger mRNA secondary structure in the start codon region. Only two of these genomes are significant for both quantities, however. Finally, a comparatively small number of genomes occupy quadrants on the right side of this scatter plot (displaying weaker aSD sequence binding strengths) and only a single phage genome is statistically different from the host for both features—in this case, having significantly weaker aSD sequence binding and stronger mRNA secondary structure in the start codon region.

All of the results that we have presented thus far treat phage genomes uniformly. However, this approach may obscure important differences in how natural selection operates on the translation-initiation regions of different phages. We thus predicted the lifestyle of each phage genome in the dataset using BACPHLIP [58] and noticed a striking difference between phages that are confidently predicted ($\geq 0.95\%$ probability) to be either temperate or virulent (n=143, of which 39 are temperate and 104 are virulent, Fig. 2B). While both categories of phages have strong aSD sequence binding energies (most data points remain to the left of zero in Fig. 2B), the virulent phages reside in the upper-left quadrant and almost exclusively have high predicted translation-initiation rates (Fig. 3B). By contrast, the vast majority of temperate phages reside in the lower-left quadrant where they display signatures of strong aSD sequence binding strengths but also stronger mRNA secondary structures as compared to the host. Repeating the full analysis from Fig. 2 without including the coding sequence GC percent and iCUB covariates reveals similar findings, with virulent phages preferentially residing in the top left quadrant, which is indicative of storng predicted translation initiation rates (Supplementary Fig. S3).

We further confirmed this stark separation between temperate and virulent phage genomes by looking at each of the features independently within the temperate and virulent phage genome sets (Supplementary Fig. S4). The clearest and most significant differences emerged when considering mRNA secondary structure around the start codon and codon usage biases. In both cases, virulent phages are predicted to have stronger translation initiation rates and translation efficiency when compared with temperate phages (Welch's t-test, $p < 0.001$). The strength of aSD sequence binding also differs slightly ($p = 0.011$) between these two sets of phages: of the 143 genomes with confident lifestyle predictions, the top 35 phages with the strongest mean aSD sequence binding energies are all virulent phages. Finally, coding sequence GC content also varied, virulent phages displaying a generally lower coding sequence GC percent ($p < 0.001$) but the link between this feature and translation efficiency is less well studied.

Finally, to better understand how these two translation initiation-related features contribute to translation efficiency—a metric that is roughly akin to the amount of protein produced from a given level of mRNA over a given time period and is often calculated using ribosome profiling. We built multi-variate regression models to predict gene-specific translation efficiencies based off of two previously published, empirically determined values [59, 60]. Both translation initiation features were highly significant, and the resulting models had Adjusted-R^2 values between 0.11–0.17 (Supplementary Fig. S5). We applied the best fitting model to predict translation efficiency values for phage genomes, and observed that a large number had significantly greater predicted mean translation efficiency when compared with the predictions for the entire *E. coli* genome. Confirming our lifestyle-based findings, this effect was driven almost entirely by virulent phage genomes—many of which had significantly greater predicted translation efficiency whereas no temperate phage genomes showed this effect (Supplementary Fig. S5).

High translation-initiation rates are a general feature of phage genomes.

We have thus far presented an in depth analysis of *E. coli* infecting phage genomes, but the generality of these findings to other host-cell bacterial species is unclear. To determine if phage mRNAs display evidence of rapid translation initiation in general, and whether this effect is particularly pronounced in virulent phages, we repeated our analyses from Fig. 1E,F on 32 additional taxonomically diverse bacterial host organisms for which we have at least 10 (dereplicated) complete phage genomes for comparison.

For nearly all host species, we found that a substantial fraction of phage genomes had significantly different aSD sequence binding strengths and mRNA secondary structure strengths compared to their hosts (Fig. 3, the fraction of significant genomes depicted in the bar chart

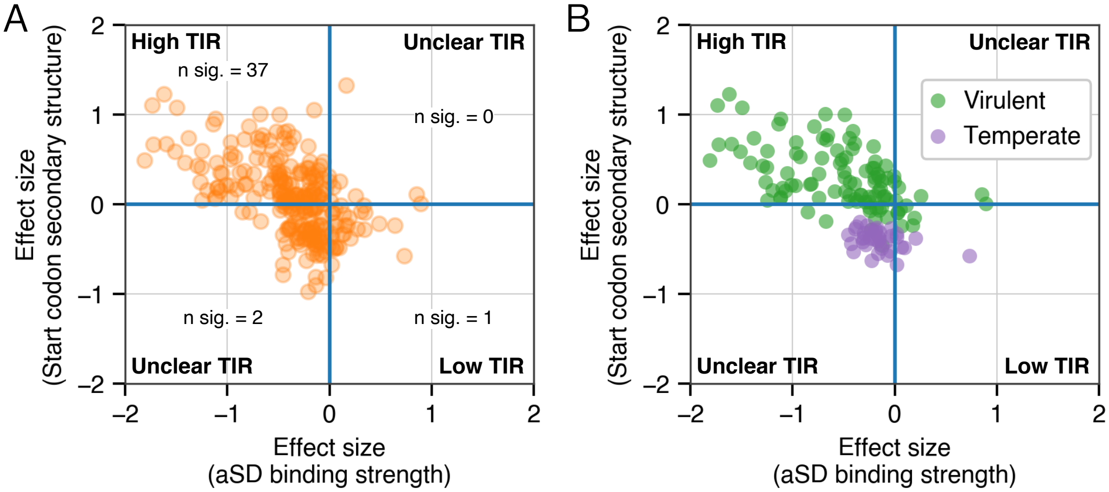


Figure 2. Multi-variable modeling of translation-initiation differences between phage and host genomes. (A) Model coefficients (standardized conditional log-odds ratio) from logistic regression comparing sequence features in translation-initiation regions between individual phage and *E. coli* genomes. Coefficients for aSD binding energies and mRNA secondary structure in the start codon region are depicted on the x- and y-axes, respectively, while coefficients for confounding variables (GC content and codon usage biases) are not shown. (B) As in (A), with phages colored according to their predicted lifestyle for the subset of phages with high-confidence lifestyle predictions ($\geq$ 0.95 probability of correct assignment using BACPHLIP).

is calculated via Welch's t-test with FDR-correction, $p < 0.01$). In nearly all cases, the effect was heavily biased in the direction of increased predicted translation-initiation rates for phage mRNAs: stronger aSD sequence binding strengths and weaker mRNA secondary structure in the start codon region. There were, however, a few scattered host species for whom there appeared to be no significant difference for one or both of the translation-initiation features. In *Cellulophaga baltica*, for instance, we note that it has previously been shown that many members of the *Bacteroidetes* phylum do not use the aSD sequence binding mechanism and it is thus unsurprising to see that phages infecting this species are also devoid of this feature [6, 61, 62]. Additionally, we observed generally weaker effects for both translation initation-related features across members of the *Firmicutes* phylum (excepting the *Bacillus* genus) but note that the translation-initiation regions in many of these host-cell genomes have particularly strong features compared with species like *E. coli*.

We additionally wished to assess differences between temperate and virulent phages for this diverse set of species. However, most phages with known, annotated lifestyles come from a biased set of host species and this bias also affects the accuracy of phage lifestyle *prediction* for different host-cell species. We predicted the lifestyle of all phages infecting the host-cell species depicted in Fig. 3 and (as in Fig 2) required a 95% probability of correct lifestyle assignment. This procedure yielded 9 host species for which we could confidently identify a minimum of 5 temperate and 5 virulent phage genomes. For these hosts, we split the phage genomes into temperate and virulent categories and repeated the analysis from Fig. 3 while

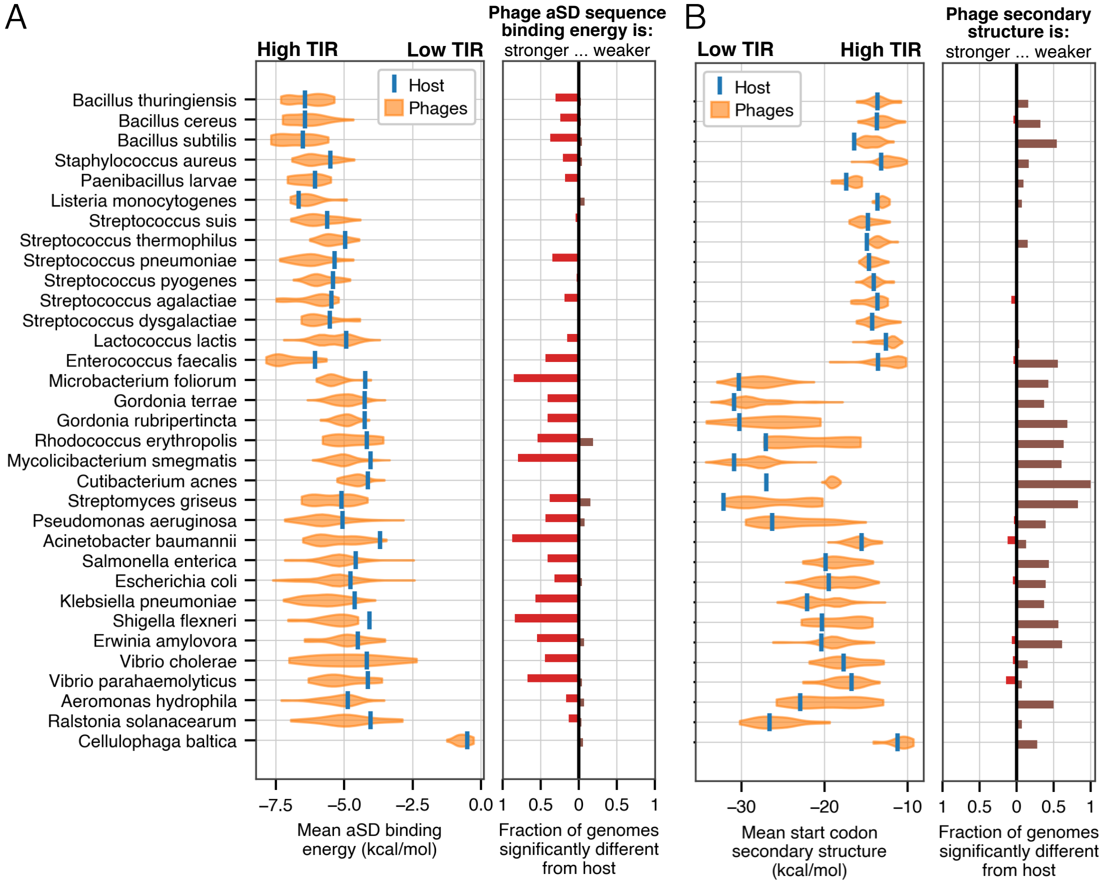


Figure 3. Phage genomes show evidence of high translation-initiation rates across a range of taxa. (A) The distribution of mean aSD binding strengths across all phages that infect an individual host is depicted as a violin plot with the host genome average shown as a blue vertical line. The fraction of phage genomes that are significantly different for each host are depicted in the vertical bar plot on the right, with red bars extending to the left of 0 indicating genomes with significantly stronger aSD binding strengths and brown bars to the right indicating significantly weaker aSD binding strengths (significance was defined via Welch's t-test with FDR-correction, $p < 0.01$)(B) Similar to (A), results for mRNA secondary structure in the start codon region. For both panels, results for *E. coli* are identical to those depicted in Fig. 1E,F.

considering temperate and virulent phages independently.

Similar to our findings across *E. coli* infecting phages, a substantial number of both virulent and temperate phages displayed stronger aSD sequence binding strengths relative to host-cell genomes and we observed comparatively minor differences in the magnitude of this feature across lifestyle classes Fig. 4A. Also confirming our results seen in *E. coli*, we observed very few cases where temperate phages were predicted to have significantly different mRNA secondary structure in the start codon region in either direction. By contrast, large numbers of virulent phages have weaker mRNA secondary structure in the start codon region Fig. 4B. Virulent phages, across nearly all of the 9 species studied, have strong sequence signatures that are indicative of rapid translation initiation. As with *E. coli*, the picture is more nuanced for temperate phages, which tend to have strong aSD sequence binding strengths (Fig. 4, left) but little or no significant difference in mRNA secondary structure strengths in the start codon region compared to host genes (Fig. 4, left).

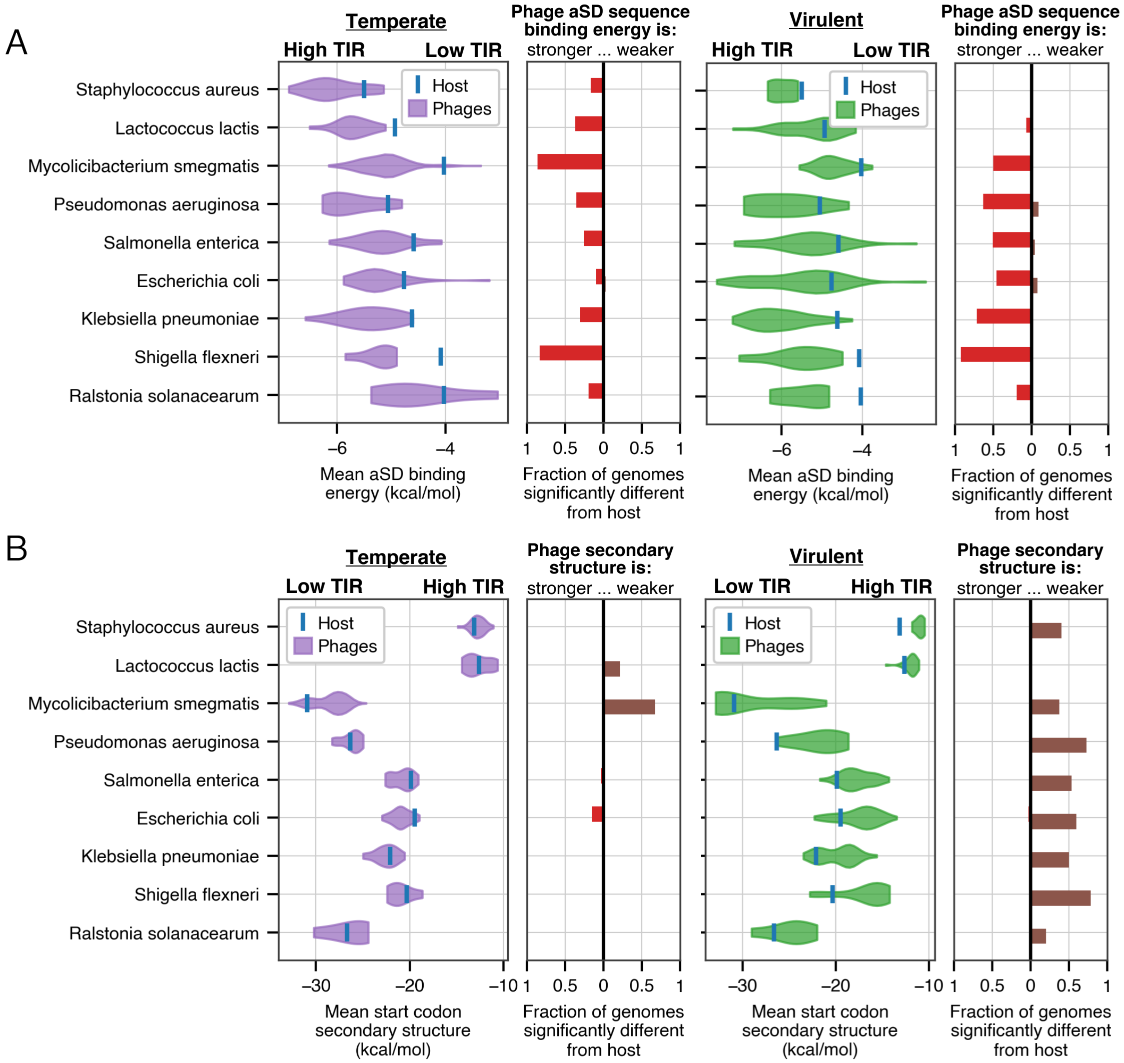


Figure 4. Phage lifestyle and translation initiation sequence feature variation across diverse bacterial species. (A) Comparing mean host-genome aSD sequence binding energies independently with values from temperate (left, purple) and virulent (right, green) phage genomes. Most virulent and temperate phage genomes across all host-cell species display strong aSD sequence binding. (B) Comparing mean host-genome secondary structure energies in the start codon region with values from temperate (left, purple) and virulent (right, green) phage genomes. Virulent phages, but not temperate phages, have weaker mRNA secondary structure in the start codon region across a range of host-species.

Discussion

Translation initiation region sequence preferences provide bacteria with a means to differentiate start codons from background sequences and to modulate the protein production rate for individual mRNAs. By analyzing the sequence patterns contained within thousands of complete bacteriophage genomes, we have shown these sequence preferences are particularly strong in phages and indicative of rapid translation-initiation rates that are on par with only the most highly expressed host-cell genes. Most strikingly, we find that virulent phage genomes differ from temperate phage genomes in this regard: mRNAs from both types of phages appear to be enriched in strong anti-Shine-Dalgarno sequence binding sites, but only virulent phages couple this ribosomal capture mechanism with particularly weak mRNA secondary structure that helps to facilitate ribosomal recruitment.

Within temperate phage genomes, the net result of partially contradictory observations—stronger aSD sequence binding and slightly stronger mRNA secondary structure relative to host genes—is unclear. To date, there have been a limited number of genome-wide transcriptomic and proteomic studies specifically targeting phage infections [42–46]. As techniques such as RNA-seq and ribosome profiling become more widespread and applied to diverse phage and host species, we expect that the details of ribosomal competition during phage infection will become more clear. However, based on regression models that were fit to empirically determined *E. coli* translation efficiency values, it appears that temperate phages are likely to have slightly lower translation efficiency (Supplementary Fig. S5). By contrast, virulent phage genomes appear to have unequivocally increased rates of translation initiation according to all of our best knowledge about this detailed molecular process. Our study thus joins a growing body of literature highlighting important differences in evolutionary processes between temperate and virulent phage genomes [63–68].

While we are unable to say precisely *why* temperate and virulent phages should differ so starkly in regards to their translational initiation regions, we speculate that the life-history of these different virus types results in unique translational pressures. Temperate phage genomes, for instance, may accumulate numerous point mutations that are likely to be under weak selection pressure during extended periods of dormancy. It is possible that excision of temperate phages from the host-cell genome (and subsequent entry into the lytic cycle) is preceded by other steps that limit competition between phage and host-cell mRNAs. By contrast, virulent phage genomes must rapidly produce protein in a race against host-cell detection mechanisms in order to reproduce at every generation. While not the focus of our study, we observed that codon usage biases also differ significantly between temperate and virulent phages, with temperate phages having generally weaker biases than we observed for virulent phage genomes. This again points in the direction of stronger overall levels of translational selection on virulent phages. However, the precise consequences of these different evolutionary modes as it relates to optimizing translation initiation remains an intriguing area of further theoretical and empirical study.

Prior work has shown that viral mRNAs (including phage) tend to have weak mRNA secondary structure surrounding the start codon, but limits in the number of available phage genomes made it difficult to explicitly study phage-host pairings [69]. Other studies have looked at differences in codon usage biases and higher-order sequence effects within bacteriophage genomes and have generally found strong signatures of translational selection in phage genomes [47–53]. Our findings thus build on existing research and suggest that a distinct an important comparison is between host and phage mRNAs, which are likely to be in direct competition for the same pool of ribosomes and translational resources during periods of active infection. The growing abundance of complete, host-linked phage genomes presents exciting opportunities for further study in this area.

A potential limitation of our work is that phylogenetic relatedness presents numerous statistical challenges. Two randomly chosen phages may share either close or distant sequence identity and common ancestry [70]. While phylogenetic comparative methods can be applied to correct for non-independence, these methods require an underlying phylogenetic tree that can be particularly challenging to assemble for phage genomes [39, 71, 72]. In lieu of this approach, we have opted for a simpler method that should partially mitigate phylogenetic biases: clustering genomes and selecting single representative genomes per cluster to ensure that the data points are more independent than they might otherwise be. Additionally, while we present multi-species comparisons between various host-cell species (Figs. 3 & 4), we again have not accounted for any confounding effects that could arise due to phylogenetic relatedness between host-cell species. However, we do not perform any explicit statistical analyses on the entire set of data, and rather include these figures to graphically show that our findings appear to apply across numerous species.

Analysis of protein coding sequences, as well as up- and down-stream regulatory sequences, has enhanced our understanding of several transcriptional and translational mechanisms, including the role of codon usage biases and higher-order sequence features in the rate of protein production [73–78]. In a sense, phages are the original synthetic biologists—exploiting host cells to produce a set of macromolecules that decrease host fitness. The precise details of how individual phages manipulate their hosts is highly varied, but higher-order genome features such as we have studied here may provide important insight into translational regulation, which may further enhance our ability to engineer both phage and host-cell genomes.

Materials and Methods

Virus data compilation, genome annotation, and host prediction

The core of our study relies on access to numerous and distinct bacteriophage genomes with consistent annotations and trustworthy predictions of primary host species. On-going research in each of these areas is continuing to expand the availability of phage genomes that are suitable for genome-linked analyses such as we performed here [39, 79–83]. To maintain consistency and to ensure access to the highest-quality genomes, we used the NCBI Virus genome database (last accessed: November 2020) and selected only "complete" genomes to include in our study. An advantage of this choice is that we additionally relied on previously existing genome annotations for all phages included in this study. Only a small subset of phage genomes have host annotations, so we limited our selection of host-cell species to those with at least 50 annotated phage genomes and 10 annotated phage genomes after dereplication (see below).

To ensure that the selected phage genomes were not severely biased by a small number of over-represented and closely related genomes, we used FastANI to measure the average nucleotide identity (ANI) between all phages that infect a single host, requiring that the alignment span a minimum of 80% the length of the shortest genome [84]. We constructed a unique all-to-all distance matrix, and used the cd-hit-est greedy algorithm to cluster phages at 95% sequence identity [85]. From each cluster, we subsequently selected the longest RefSeq genome as a cluster representative. In the event that there were no RefSeq genomes in a cluster, we simply selected the longest genome. Finally, we removed poorly annotated phage genomes from the analysis, which we conservatively defined as those with an annotated coding sequence density <50%. For *E. coli*, our processing began with a set of 1,473 genomes and our final dataset encompassed 254 genomes, which were analyzed throughout this manuscript.

Definitions of sequence features related to translation initiation.

The Shine-Dalgarno sequence is a collection of related mRNA sequence motifs that are defined by their ability to bind strongly to the highly conserved anti-Shine-Dalgarno sequence—located on the 30S subunit of the bacterial ribosome. Here, we define the aSD binding strength of a given coding sequence as the strongest possible binding affinity between the anti-SD sequence (defined via the core sequence of 5′-CCUCCU-3′) and hexamer sequences upstream of the start codon, restricting the upstream range to have a minimum gap of 4 nucleotides and a maximum gap of 10 nucleotides between the 3′ most base in the hexamer and the first nucleotide of the start codon. Binding energies were calculated using the "RNAcofold" program, part of the ViennaRNA (v2.4.14) suite [86].

We determined the strength of mRNA secondary structure surrounding all start codons by assessing a 90 nucleotide long sequence fragment (30 nts upstream of the start codon and 60 nts downstream) using the "RNAfold" program from ViennaRNA (v2.4.14) [86]. While we expect this to be a rough approximation of the true secondary structure present in the start codon region for each mRNA, numerous prior studies have showed that similar window sizes roughly capture the relevant feature of mRNA secondary structure strength [22, 56, 57].

More complicated models of translation initiation may include explicit penalties for SD sequence spacing relative to the start codon [23], the identity of the start codon itself [87], kinetics of secondary structure unfolding and refolding [88], penalties for binding too strongly to the aSD sequence [56], as well as a number of other possible features [89–91]. However, it is unknown how transferable these other mechanisms are across diverse organisms given that the vast majority of existing work has been performed in *E. coli*. To keep our model simple, we focused on the two most consistent and frequently cited modifiers of translation-initiation rates. Additionally, we found that a simple multi-variable linear regression model consisting of only aSD binding strength and the strength of secondary structure around the start codon is capable of significantly predicting translation efficiencies derived from two separate genome-wide *E. coli* datasets [59, 60] (Supplementary Fig. S5). In these models, R^2

values ranged from 0.11–0.17 and predictions of both models were highly correlated. Further, in both cases, start codons with weak mRNA secondary structure and strong aSD sequence binding have the highest empirically measured translation efficiency values and both sequence features were highly significant in the regression models ($p < 0.001$). We did not apply these models to predict translation efficiencies for phages or hosts outside of *E. coli*, reasoning that doing so would make a dangerous assumption that the mechanisms of translational regulation remain similar across phylogenetically diverse species. Our current approach does, however, assume that the mechanisms of translation efficiency and translational regulation remain similar between normal *E. coli* growth and during phage infection.

Other CDS-level feature definitions and inclusion criteria

Codon usage bias is frequently considered an indicator of translational selection due to the potential for synonymous codons to modulate the rate of translation elongation. We thus wanted to ensure that our findings were robust to variation in codon usage bias differences between host and phage genomes, as well as GC content variation (which has a direct impact on mRNA secondary structure strength). Coding sequence GC content was simply calculated for each CDS as the number of G+C nucleotides divided by the total coding sequence length. Codon usage bias was measured using iCUB [92]. There are many possible metrics of codon usage bias, but a benchmark study showed that iCUB outperformed other metrics that rely solely on coding sequence information (as opposed to *a priori* defined reference sets of genes or other external information such as tRNA abundances or a reference set of highly expressed genes). The iCUB metric is similar to the more commonly cited effective number of codons: lower values indicate fewer codons and thus more bias. However, iCUB explicitly controls for GC content variation and produces better estimates of protein abundance across diverse microbial taxa. We observed that both of these features (GC-content and iCUB) differ significantly between phage and host species (Supplementary Fig. S2), providing validation that our results are likely to be more robust and conservative by controlling for these effects.

For all studied coding sequences (either phage or host-derived), we only considered annotated genes whose length was at least 90 nucleotides (30 amino acids) and for which the length was a multiple of three (potentially removing a small number of genes with programmed frame-shifts). Additionally, all of our results for host genomes excluded coding sequences whose "product" feature annotation contained the word "phage". This filter was performed to ideally exclude prophage-associated coding sequences from being included in the host genome when drawing comparisons.

Compilation of *E. coli*-specific empirical data

We leveraged several existing genome-scale, empirical data sources to thoroughly contrast *E. coli* infecting phages with various subsets of *E. coli* genes as well as to ensure that our translation-initiation metrics were associated with empirically derived translation efficiency measurements. Specifically, we relied on PAXdb for protein abundance data [93], two separate datasets of gene essentiality [94, 95], and two separate datasets of ribosome-profiling derived translation efficiency measurements (briefly referenced in the preceding section) [59, 60].

For protein abundance data, we simply drew various percentile-based thresholds to consider increasingly stringent sets of "highly-expressed" genes. PAXdb aggregates protein abundance measurements from numerous studies and growth conditions, and while these abundance values are estimates from only a small snap-shot of possible studies and possible growth conditions, these values are well-established (if rough) indicators of overall protein abundance [93]. For gene essentiality, we used a consensus approach where we only considered genes that were considered either "essential" or "non-essential" in two separate datasets (to increase robustness) and did not consider genes with conflicting annotations for this portion of the analysis.

Phage lifestyle prediction

We used the BACPHLIP software [58] to categorize all phage genomes in our dataset into temperate and virulent lifestyles. BACPHLIP uses genome-sequence input, determines the presence/absence of several hundred lysogeny-associated protein domains, and uses a random forest classifier return a probability of the given phage being either temperate or virulent. Here, we limited our analyses of both *E. coli* infecting phages and phages that infect other species to those with a predicted lifestyle probability ≥ 0.95. Additionally, when assessing differences between temperate and virulent phages, we removed any host species from our analysis that did not have at least 5 phage genomes (after dereplication) in each lifestyle category.

Statistical tests

Single-variable comparisons were performed in Python using the scipy package implementation of Welch's T-test. Multiple comparisons were corrected for by using the applying the statsmodels implementation of FDR correction (Benjamini/Hochberg) to the list of $p-$values with alpha set to 0.01. Multivariate analyses were performed using the logit function in statsmodels with zscore normalization to all predictor variables in order to standardize effect sizes. Correction for multiple comparisons in these models was again accomplished by filtering the $p-$values from individual model coefficients through the FDR correction procedure with the alpha value set to 0.01. Thus, we ensured that all of our results were highly significant and robust to artifacts arising from multiple comparisons.

Host taxonomy assignment

We did not explicitly model any effects across a phylogenetic tree and instead present analyses of multiple species independently in Figs. 3 & 4. We do, however, note that no phage genomes were shared across host species for any part of this analysis, but individual host species nevertheless have their own complicated and non-independent phylogenetic histories. We arranged species taxonomically for easier visual comparison by querying the NCBI taxonomy database and used the ete3 package to roughly order species according to their taxonomic grouping. All statistical results were performed for each species independently.

Code and data availability.

All code and data necessary to re-create the analyses in this manuscript are available at `https://github.com/adamhockenberry/phage-translation` and `https://doi.org/10.5281/zenodo.4708008`, respectively.

Acknowledgements

This work was supported by National Institutes of Health grants R01 GM088344 to C.O.W. and F32 GM130113 to A.J.H. C.O.W. also received support from the Jane and Roland Blumberg Centennial Professorship in Molecular Evolution and the Dwight W. and Blanche Faye Reeder Centennial Fellowship in Systematic and Evolutionary Biology at The University of Texas at Austin. The Texas Advanced Computing Center (TACC) at The University of Texas at Austin provided high-performance computing resources. The authors additionally acknowledge valuable feedback and support from members of the Wilke lab.

Competing financial interest

The authors have declared that no competing interests exist.

References

[1] Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F. & Sockett, R. E. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Research* **33**, 1141–1153 (2005). URL `https://academic.oup.com/nar/article/33/4/1141/1521122`. Publisher: Oxford Academic.

[2] Dekel, E. & Alon, U. Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436**, 588–592 (2005).

[3] Vieira-Silva, S. & Rocha, E. P. C. The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics. *PLOS Genetics* **6**, e1000808 (2010). URL `https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000808`. Publisher: Public Library of Science.

[4] Botzman, M. & Margalit, H. Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome Biology* **12**, R109 (2011). URL `https://doi.org/10.1186/gb-2011-12-10-r109`.

[5] Eames, M. & Kortemme, T. Cost-benefit tradeoffs in engineered lac operons. *Science* **336**, 911–915 (2012).

[6] Hockenberry, A. J., Stern, A. J., Amaral, L. A. N. & Jewett, M. C. Diversity of Translation Initiation Mechanisms across Bacterial Species Is Driven by Environmental Conditions and Growth Demands. *Molecular Biology and Evolution* **35**, 582–592 (2018).

URL https://academic.oup.com/mbe/article/35/3/582/4705836. Publisher: Oxford Academic.

[7] LaBella, A. L., Opulente, D. A., Steenwyk, J. L., Hittinger, C. T. & Rokas, A. Variation and selection on codon usage bias across an entire subphylum. *PLOS Genetics* **15**, e1008304 (2019). URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008304. Publisher: Public Library of Science.

[8] Hausser, J., Mayo, A., Keren, L. & Alon, U. Central dogma rates and the trade-off between precision and economy in gene expression. *Nature Communications* **10**, 68 (2019).

[9] Burge, C., Campbell, A. M. & Karlin, S. Over- and under-representation of short oligonucleotides in DNA sequences. *Proceedings of the National Academy of Sciences* **89**, 1358–1362 (1992). URL https://www.pnas.org/content/89/4/1358. Publisher: National Academy of Sciences Section: Research Article.

[10] Hahn, M. W., Stajich, J. E. & Wray, G. A. The Effects of Selection Against Spurious Transcription Factor Binding Sites. *Molecular Biology and Evolution* **20**, 901–906 (2003). URL https://academic.oup.com/mbe/article/20/6/901/982691. Publisher: Oxford Academic.

[11] Froula, J. L. & Francino, M. P. Selection against Spurious Promoter Motifs Correlates with Translational Efficiency across Bacteria. *PLOS ONE* **2**, e745 (2007). URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0000745. Publisher: Public Library of Science.

[12] Tats, A., Tenson, T. & Remm, M. Preferred and avoided codon pairs in three domains of life. *BMC Genomics* **9**, 463 (2008). URL https://doi.org/10.1186/1471-2164-9-463.

[13] Itzkovitz, S., Hodis, E. & Segal, E. Overlapping codes within protein-coding sequences. *Genome Research* **20**, 1582–1589 (2010). URL http://genome.cshlp.org/content/20/11/1582. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[14] Villada, J. C., Duran, M. F. & Lee, P. K. H. Interplay between Position-Dependent Codon Usage Bias and Hydrogen Bonding at the 5' End of ORFeomes. *mSystems* **5** (2020). URL https://msystems.asm.org/content/5/4/e00613-20. Publisher: American Society for Microbiology Journals Section: Research Article.

[15] Shine, J. & Dalgarno, L. The 3'-Terminal Sequence of Escherichia coli 16S Ribosomal RNA: Complementarity to Nonsense Triplets and Ribosome Binding Sites. *Proceedings of the National Academy of Sciences* **71**, 1342–1346 (1974). URL https://www.pnas.org/content/71/4/1342. Publisher: National Academy of Sciences Section: Biological Sciences: Biochemistry.

[16] Shine, J. & Dalgarno, L. Determinant of cistron specificity in bacterial ribosomes. *Na-*

ture **254**, 34–38 (1975). URL https://www.nature.com/articles/254034a0. Number: 5495 Publisher: Nature Publishing Group.

[17] Nakagawa, S., Niimura, Y., Miura, K.-i. & Gojobori, T. Dynamic evolution of translation initiation mechanisms in prokaryotes. *Proceedings of the National Academy of Sciences* **107**, 6382–6387 (2010). URL https://www.pnas.org/content/107/14/6382. Publisher: National Academy of Sciences Section: Biological Sciences.

[18] Omotajo, D., Tate, T., Cho, H. & Choudhary, M. Distribution and diversity of ribosome binding sites in prokaryotic genomes. *BMC Genomics* **16**, 604 (2015). URL https://doi.org/10.1186/s12864-015-1808-6.

[19] Yang, C., Hockenberry, A. J., Jewett, M. C. & Amaral, L. A. N. Depletion of Shine-Dalgarno Sequences Within Bacterial Coding Regions Is Expression Dependent. *G3: Genes, Genomes, Genetics* **6**, 3467–3474 (2016). URL https://www.g3journal.org/content/6/11/3467. Publisher: G3: Genes, Genomes, Genetics Section: Investigations.

[20] Diwan, G. D. & Agashe, D. The Frequency of Internal Shine–Dalgarno-like Motifs in Prokaryotes. *Genome Biology and Evolution* **8**, 1722–1733 (2016). URL https://academic.oup.com/gbe/article/8/6/1722/2574012. Publisher: Oxford Academic.

[21] Hockenberry, A. J., Jewett, M. C., Amaral, L. A. N. & Wilke, C. O. Within-Gene Shine–Dalgarno Sequences Are Not Selected for Function. *Molecular Biology and Evolution* **35**, 2487–2498 (2018). URL https://academic.oup.com/mbe/article/35/10/2487/5063446. Publisher: Oxford Academic.

[22] Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-Sequence Determinants of Gene Expression in Escherichia coli. *Science* **324**, 255–258 (2009). URL https://science.sciencemag.org/content/324/5924/255. Publisher: American Association for the Advancement of Science Section: Report.

[23] Salis, H. M., Mirsky, E. A. & Voigt, C. A. Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology* **27**, 946–950 (2009). URL https://www.nature.com/articles/nbt.1568. Number: 10 Publisher: Nature Publishing Group.

[24] Tuller, T. *et al.* An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. *Cell* **141**, 344–354 (2010). URL http://www.sciencedirect.com/science/article/pii/S0092867410003193.

[25] Qian, W., Yang, J.-R., Pearson, N. M., Maclean, C. & Zhang, J. Balanced Codon Usage Optimizes Eukaryotic Translational Efficiency. *PLOS Genetics* **8**, e1002603 (2012). URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002603. Publisher: Public Library of Science.

[26] Pechmann, S. & Frydman, J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nature Structural & Molecular Biology* **20**, 237–243 (2013). URL https://www.nature.com/articles/nsmb.2466. Number: 2 Publisher: Nature Publishing Group.

[27] Qian, L. & Kussell, E. Genome-Wide Motif Statistics are Shaped by DNA Binding Proteins over Evolutionary Time Scales. *Physical Review X* **6**, 041009 (2016). URL `https://link.aps.org/doi/10.1103/PhysRevX.6.041009`. Publisher: American Physical Society.

[28] Umu, S. U., Poole, A. M., Dobson, R. C. & Gardner, P. P. Avoidance of stochastic RNA interactions can be harnessed to control protein expression levels in bacteria and archaea. *eLife* **5**, e13479 (2016). URL `https://doi.org/10.7554/eLife.13479`. Publisher: eLife Sciences Publications, Ltd.

[29] Chaney, J. L. *et al.* Widespread position-specific conservation of synonymous rare codons within coding sequences. *PLOS Computational Biology* **13**, e1005531 (2017). URL `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005531`. Publisher: Public Library of Science.

[30] Chen, S. *et al.* Codon-Resolution Analysis Reveals a Direct and Context-Dependent Impact of Individual Synonymous Mutations on mRNA Level. *Molecular Biology and Evolution* **34**, 2944–2958 (2017). URL `https://academic.oup.com/mbe/article/34/11/2944/4093241`. Publisher: Oxford Academic.

[31] Jacobs, W. M. & Shakhnovich, E. I. Evidence of evolutionary selection for cotranslational folding. *Proceedings of the National Academy of Sciences* **114**, 11434–11439 (2017). URL `https://www.pnas.org/content/114/43/11434`. Publisher: National Academy of Sciences Section: Biological Sciences.

[32] Cambray, G., Guimaraes, J. C. & Arkin, A. P. Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in Escherichia coli. *Nature Biotechnology* **36**, 1005–1015 (2018). URL `https://www.nature.com/articles/nbt.4238`. Number: 10 Publisher: Nature Publishing Group.

[33] Deng, L. *et al.* Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. *Nature* **513**, 242–245 (2014). URL `http://www.nature.com/articles/nature13459`.

[34] Roux, S., Hallam, S. J., Woyke, T. & Sullivan, M. B. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife* **4**, e08490 (2015).

[35] Paez-Espino, D. *et al.* Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).

[36] Nishimura, Y. *et al.* Environmental Viral Genomes Shed New Light on Virus-Host Interactions in the Ocean. *mSphere* **2**, e00359–16 (2017). URL `https://mSphere.asm.org/lookup/doi/10.1128/mSphere.00359-16`.

[37] Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nature Microbiology* **3**, 870–880 (2018). URL `http://www.nature.com/articles/s41564-018-0190-y`.

[38] Tisza, M. J. *et al.* Discovery of several thousand highly diverse circular DNA viruses. *eLife* **9**, e51971 (2020). URL `https://doi.org/10.7554/eLife.51971`.

[39] Roux, S. *et al.* IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Research* **49**, D764–D775 (2021). URL `https://doi.org/10.1093/nar/gkaa946`.

[40] Ofir, G. & Sorek, R. Contemporary Phage Biology: From Classic Models to New Insights. *Cell* **172**, 1260–1270 (2018).

[41] Dion, M. B., Oechslin, F. & Moineau, S. Phage diversity, genomics and phylogeny. *Nature Reviews Microbiology* **18**, 125–138 (2020).

[42] Liu, X., Jiang, H., Gu, Z. & Roberts, J. W. High-resolution view of bacteriophage lambda gene expression by ribosome profiling. *Proceedings of the National Academy of Sciences* **110**, 11928–11933 (2013). URL `https://www.pnas.org/content/110/29/11928`. ISBN: 9781309739112 Publisher: National Academy of Sciences Section: Biological Sciences.

[43] Jack, B. R. *et al.* Reduced protein expression in a virus attenuated by codon deoptimization. *G3: Genes, Genomes, Genetics* **7**, 2957–2968 (2017).

[44] Jack, B. R., Boutz, D. R., Paff, M. L., Smith, B. L. & Wilke, C. O. Transcript degradation and codon usage regulate gene expression in a lytic phage. *Virus Evolution* **5** (2019).

[45] Jaschke, P. R., Dotson, G. A., Hung, K. S., Liu, D. & Endy, D. Definitive demonstration by synthesis of genome annotation completeness. *Proceedings of the National Academy of Sciences* **116**, 24206–24213 (2019).

[46] Logel, D. Y. & Jaschke, P. R. A high-resolution map of bacteriophage phiX174 transcription. *Virology* **547**, 47–56 (2020).

[47] Xia, X. & Yuen, K. Y. Differential selection and mutation between dsDNA and ssDNA phages shape the evolution of their genomic AT percentage. *BMC Genetics* **6**, 20 (2005). URL `https://doi.org/10.1186/1471-2156-6-20`.

[48] Carbone, A. Codon Bias is a Major Factor Explaining Phage Evolution in Translationally Biased Hosts. *Journal of Molecular Evolution* **66**, 210–223 (2008). URL `https://doi.org/10.1007/s00239-008-9068-6`.

[49] Lucks, J. B., Nelson, D. R., Kudla, G. R. & Plotkin, J. B. Genome Landscapes and Bacteriophage Codon Usage. *PLOS Computational Biology* **4**, e1000001 (2008). URL `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000001`. Publisher: Public Library of Science.

[50] Chithambaram, S., Prabhakaran, R. & Xia, X. Differential Codon Adaptation between dsDNA and ssDNA Phages in Escherichia coli. *Molecular Biology and Evolution* **31**, 1606–1617 (2014). URL `https://doi.org/10.1093/molbev/msu087`.

[51] Chithambaram, S., Prabhakaran, R. & Xia, X. The Effect of Mutation and Selection on Codon Adaptation in Escherichia coli Bacteriophage. *Genetics* **197**, 301–315 (2014). URL `https://www.genetics.org/content/197/1/301`. Publisher: Genetics Section: Investigations.

[52] Mioduser, O., Goz, E. & Tuller, T. Significant differences in terms of codon usage bias between bacteriophage early and late genes: a comparative genomics analysis. *BMC Genomics* **18**, 866 (2017). URL `https://doi.org/10.1186/s12864-017-4248-7`.

[53] Goz, E., Zafrir, Z. & Tuller, T. Universal evolutionary selection for high dimensional silent patterns of information hidden in the redundancy of viral genetic code. *Bioinformatics* **34**, 3241–3248 (2018). URL `https://academic.oup.com/bioinformatics/article/34/19/3241/4990489`. Publisher: Oxford Academic.

[54] Dedrick, R. M. *et al.* Engineered bacteriophages for treatment of a patient with a disseminated drug-resistant Mycobacterium abscessus. *Nature Medicine* **25**, 730–733 (2019).

[55] Lammens, E.-M., Nikel, P. I. & Lavigne, R. Exploring the synthetic biology potential of bacteriophages for engineering non-model bacteria. *Nature Communications* **11**, 5294 (2020).

[56] Hockenberry, A. J., Pah, A. R., Jewett, M. C. & Amaral, L. A. N. Leveraging genome-wide datasets to quantify the functional role of the anti-Shine–Dalgarno sequence in regulating translation efficiency. *Open Biology* **7**, 160239 (2017). URL `https://royalsocietypublishing.org/doi/full/10.1098/rsob.160239`. Publisher: Royal Society.

[57] Terai, G. & Asai, K. Improving the prediction accuracy of protein abundance in Escherichia coli using mRNA accessibility. *Nucleic Acids Research* gkaa481 (2020). URL `https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkaa481/5854142`.

[58] Hockenberry, A. J. & Wilke, C. O. BACPHLIP: Predicting bacteriophage lifestyle from conserved protein domains. *bioRxiv* 2020.05.13.094805 (2020). URL `https://www.biorxiv.org/content/10.1101/2020.05.13.094805v1`. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[59] Li, G.-W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. *Cell* **157**, 624–635 (2014). URL `https://www.cell.com/cell/abstract/S0092-8674(14)00232-3`. Publisher: Elsevier.

[60] Gorochowski, T. E. *et al.* Absolute quantification of translational regulation and burden using combined sequencing approaches. *Molecular Systems Biology* **15**, e8719 (2019). URL `https://www.embopress.org/doi/full/10.15252/msb.20188719`. Publisher: John Wiley & Sons, Ltd.

[61] Baez, W. D. *et al.* Global analysis of protein synthesis in Flavobacterium johnsoniae reveals the use of Kozak-like sequences in diverse bacteria. *Nucleic Acids Research* **47**, 10477–10488 (2019). URL `https://doi.org/10.1093/nar/gkz855`.

[62] Jha, V. *et al.* Structural basis of sequestration of the anti-Shine-Dalgarno sequence in the Bacteroidetes ribosome. *Nucleic Acids Research* **49**, 547–567 (2021). URL `https://doi.org/10.1093/nar/gkaa1195`.

[63] Bobay, L.-M., Rocha, E. P. & Touchon, M. The Adaptation of Temperate Bacteriophages to Their Host Genomes. *Molecular Biology and Evolution* **30**, 737–751 (2013). URL `https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/mss279`.

[64] Bobay, L.-M., Touchon, M. & Rocha, E. P. C. Pervasive domestication of defective prophages by bacteria. *Proceedings of the National Academy of Sciences* **111**, 12127–12132 (2014). URL `https://www.pnas.org/content/111/33/12127`. Publisher: National Academy of Sciences Section: Biological Sciences.

[65] Touchon, M., Bernheim, A. & Rocha, E. P. Genetic and life-history traits associated with the distribution of prophages in bacteria. *The ISME Journal* **10**, 2744–2754 (2016). URL `http://www.nature.com/articles/ismej201647`.

[66] Mavrich, T. N. & Hatfull, G. F. Bacteriophage evolution differs by host, lifestyle and genome. *Nature Microbiology* **2**, 17112 (2017). URL `http://www.nature.com/articles/nmicrobiol2017112`.

[67] Howard-Varona, C., Hargreaves, K. R., Abedon, S. T. & Sullivan, M. B. Lysogeny in nature: mechanisms, impact and ecology of temperate phages. *The ISME Journal* **11**, 1511–1520 (2017). URL `https://www.nature.com/articles/ismej201716`. Number: 7 Publisher: Nature Publishing Group.

[68] Sousa, J. A. M. d., Pfeifer, E., Touchon, M. & Rocha, E. P. C. Causes and consequences of bacteriophage diversification via genetic exchanges across lifestyles and bacterial taxa. *bioRxiv* 2020.04.14.041137 (2021). URL `https://www.biorxiv.org/content/10.1101/2020.04.14.041137v2`. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[69] Zhou, T. & Wilke, C. O. Reduced stability of mRNA secondary structure near the translation-initiation site in dsDNA viruses. *BMC Evolutionary Biology* **11**, 59 (2011). URL `https://doi.org/10.1186/1471-2148-11-59`.

[70] Bobay, L.-M. & Ochman, H. Biological species in the viral world. *Proceedings of the National Academy of Sciences* **115**, 6040–6045 (2018). URL `http://www.pnas.org/lookup/doi/10.1073/pnas.1717593115`.

[71] Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nature Biotechnology* **37**, 632–639 (2019). URL `https://www.nature.com/articles/s41587-019-0100-8`. Number: 6 Publisher: Nature Publishing Group.

[72] Low, S. J., Džunková, M., Chaumeil, P.-A., Parks, D. H. & Hugenholtz, P. Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order Caudovirales. *Nature Microbiology* **4**, 1306–1315 (2019). URL `https://www.nature.com/articles/s41564-019-0448-z`. Number: 8 Publisher: Nature Publishing Group.

[73] Clarke IV, T. F. & Clark, P. L. Rare Codons Cluster. *PLOS ONE* **3**, e3412

(2008). URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0003412`. Publisher: Public Library of Science.

[74] Gu, W., Zhou, T. & Wilke, C. O. A Universal Trend of Reduced mRNA Stability near the Translation-Initiation Site in Prokaryotes and Eukaryotes. *PLOS Computational Biology* **6**, e1000664 (2010). URL `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000664`. Publisher: Public Library of Science.

[75] Hockenberry, A. J., Sirer, M. I., Amaral, L. A. N. & Jewett, M. C. Quantifying Position-Dependent Codon Usage Bias. *Molecular Biology and Evolution* **31**, 1880–1893 (2014). URL `https://academic.oup.com/mbe/article/31/7/1880/2925740`. Publisher: Oxford Academic.

[76] Zur, H. & Tuller, T. Exploiting hidden information interleaved in the redundancy of the genetic code without prior knowledge. *Bioinformatics* **31**, 1161–1168 (2015). URL `https://academic.oup.com/bioinformatics/article/31/8/1161/212401`. Publisher: Oxford Academic.

[77] Novoa, E. M., Jungreis, I., Jaillon, O. & Kellis, M. Elucidation of Codon Usage Signatures across the Domains of Life. *Molecular Biology and Evolution* **36**, 2328–2339 (2019). URL `https://academic.oup.com/mbe/article/36/10/2328/5492082`. Publisher: Oxford Academic.

[78] Deng, Y., de Lima Hedayioglu, F., Kalfon, J., Chu, D. & von der Haar, T. Hidden patterns of codon usage bias across kingdoms. *Journal of The Royal Society Interface* **17**, 20190819 (2020). URL `https://royalsocietypublishing.org/doi/full/10.1098/rsif.2019.0819`. Publisher: Royal Society.

[79] Galiez, C., Siebert, M., Enault, F., Vincent, J. & Söding, J. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* **33**, 3113–3114 (2017). URL `https://academic.oup.com/bioinformatics/article/33/19/3113/3964377`. Publisher: Oxford Academic.

[80] Gao, N. L. *et al.* MVP: a microbe–phage interaction database. *Nucleic Acids Research* **46**, D700–D707 (2018). URL `https://doi.org/10.1093/nar/gkx1124`.

[81] Ecale Zhou, C. L. *et al.* multiPhATE: bioinformatics pipeline for functional annotation of phage isolates. *Bioinformatics* **35**, 4402–4404 (2019). URL `https://academic.oup.com/bioinformatics/article/35/21/4402/5488969`. Publisher: Oxford Academic.

[82] McNair, K., Zhou, C., Dinsdale, E. A., Souza, B. & Edwards, R. A. PHANOTATE: a novel approach to gene identification in phage genomes. *Bioinformatics* **35**, 4537–4542 (2019). URL `https://academic.oup.com/bioinformatics/article/35/22/4537/5480131`. Publisher: Oxford Academic.

[83] Young, F., Rogers, S. & Robertson, D. L. Predicting host taxonomic information from viral genomes: A comparison of feature representations. *PLOS Computational Biology* **16**, e1007894 (2020). URL `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007894`. Publisher: Public Library of Science.

[84] Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* **9**, 5114 (2018). URL `http://www.nature.com/articles/s41467-018-07641-9`.

[85] Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012). URL `https://doi.org/10.1093/bioinformatics/bts565`.

[86] Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms for Molecular Biology* **6**, 26 (2011). URL `https://doi.org/10.1186/1748-7188-6-26`.

[87] Hecht, A. *et al.* Measurements of translation initiation from all 64 codons in E. coli. *Nucleic Acids Research* **45**, 3615–3626 (2017). URL `https://academic.oup.com/nar/article/45/7/3615/2990259`. Publisher: Oxford Academic.

[88] Espah Borujeni, A. & Salis, H. M. Translation Initiation is Controlled by RNA Folding Kinetics via a Ribosome Drafting Mechanism. *Journal of the American Chemical Society* **138**, 7016–7023 (2016). URL `https://doi.org/10.1021/jacs.6b01453`. Publisher: American Chemical Society.

[89] Verma, M. *et al.* A short translational ramp determines the efficiency of protein synthesis. *Nature Communications* **10**, 5774 (2019). URL `https://www.nature.com/articles/s41467-019-13810-1`. Number: 1 Publisher: Nature Publishing Group.

[90] Osterman, I. A. *et al.* Translation at first sight: the influence of leading codons. *Nucleic Acids Research* gkaa430 (2020). URL `https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkaa430/5840581`.

[91] Kuo, S.-T. *et al.* Global fitness landscapes of the Shine-Dalgarno sequence. *Genome Research* (2020). URL `http://genome.cshlp.org/content/early/2020/05/18/gr.260182.119`. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[92] Liu, S. S., Hockenberry, A. J., Jewett, M. C. & Amaral, L. A. N. A novel framework for evaluating the performance of codon usage bias metrics. *Journal of The Royal Society Interface* **15**, 20170667 (2018). URL `https://royalsocietypublishing.org/doi/full/10.1098/rsif.2017.0667`. Publisher: Royal Society.

[93] Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D. & Mering, C. v. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *PROTEOMICS* **15**, 3163–3168 (2015). URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.201400441`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pmic.201400441.

[94] Goodall, E. C. A. *et al.* The Essential Genome of Escherichia coli K-12. *mBio* **9** (2018). URL `https://mbio.asm.org/content/9/1/e02096-17`. Publisher: American Society for Microbiology Section: Research Article.

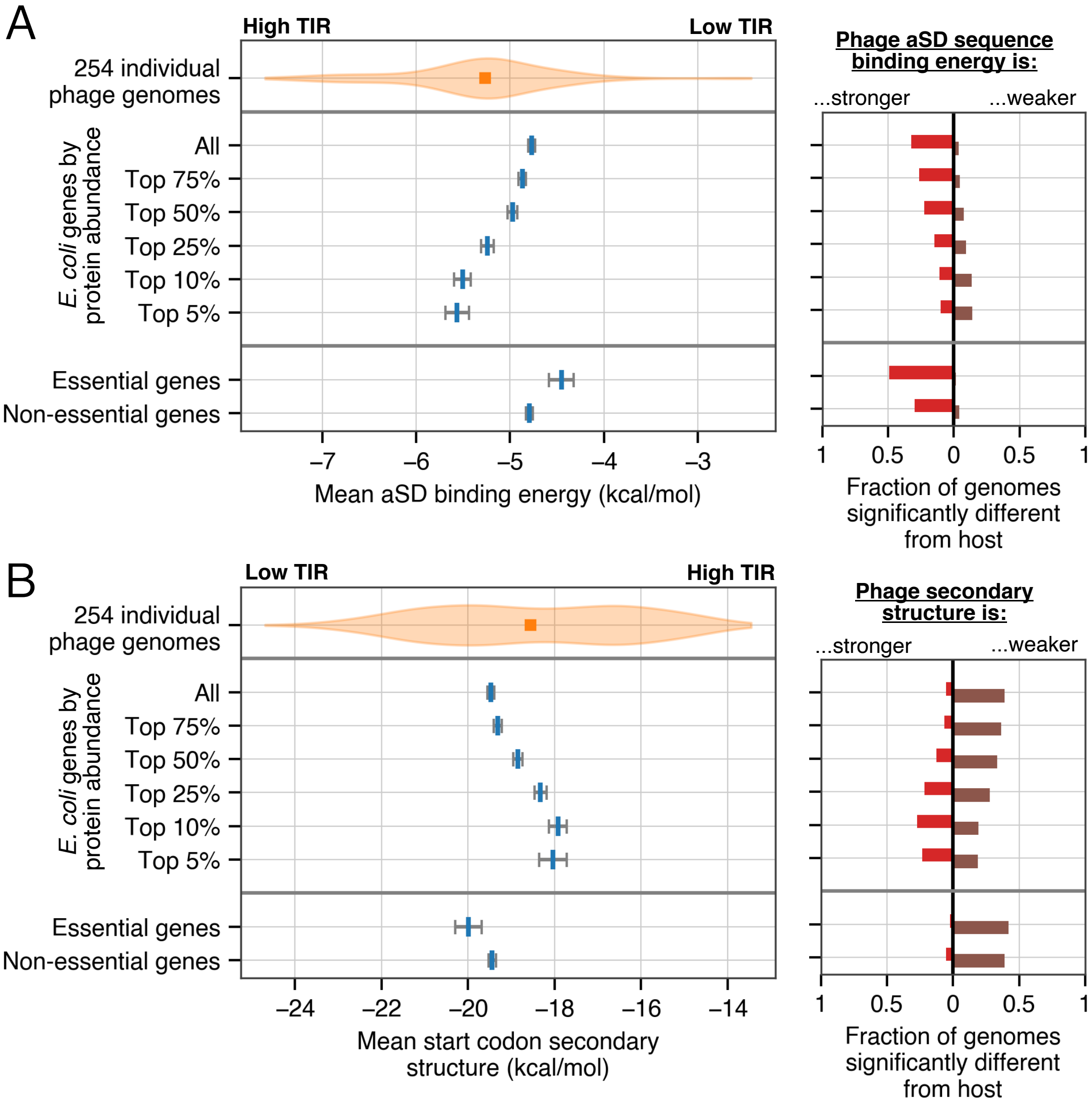
[95] Price, M. N. *et al.* Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* **557**, 503–509 (2018). URL `https://www.nature.com/articles/s41586-018-0124-0`. Number: 7706 Publisher: Nature Publishing Group.

Supplementary Information: Rapid translation initiation is a general feature of virulent phage genomes.

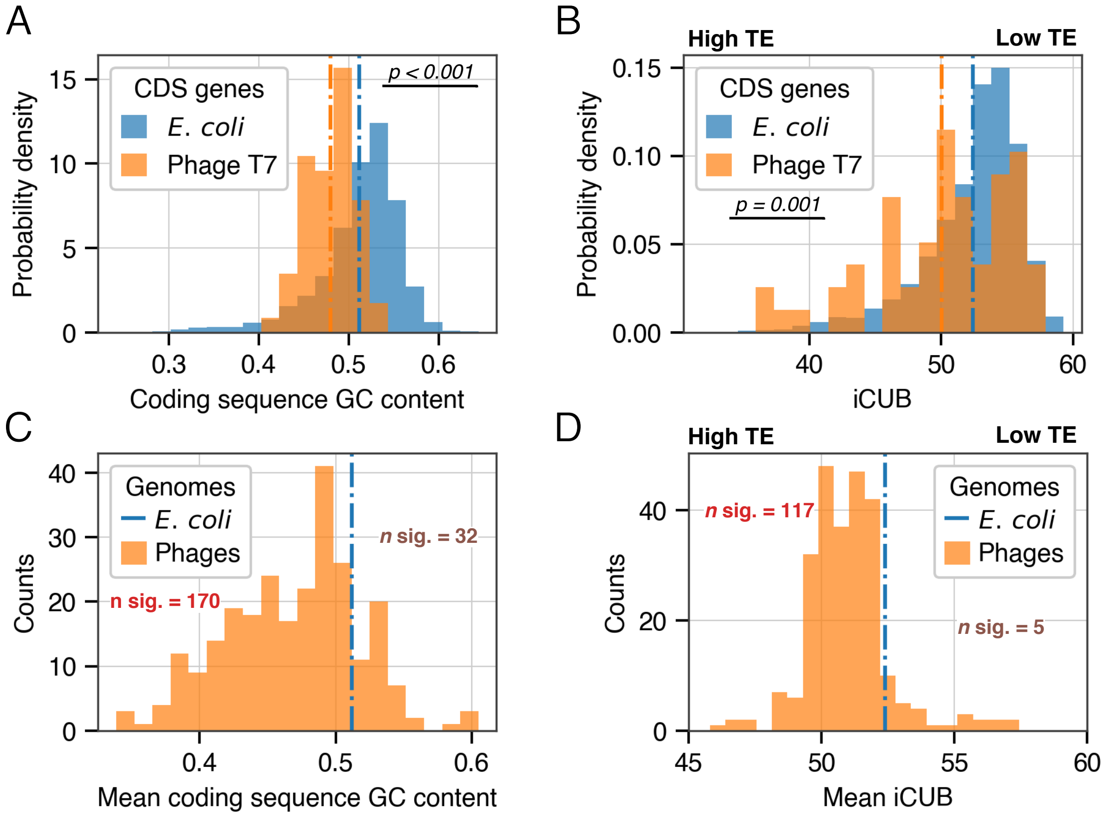
Adam J. Hockenberry[1,*], David C. Weaver[1], Claus O. Wilke[1]

1 Department of Integrative Biology, The University of Texas at Austin

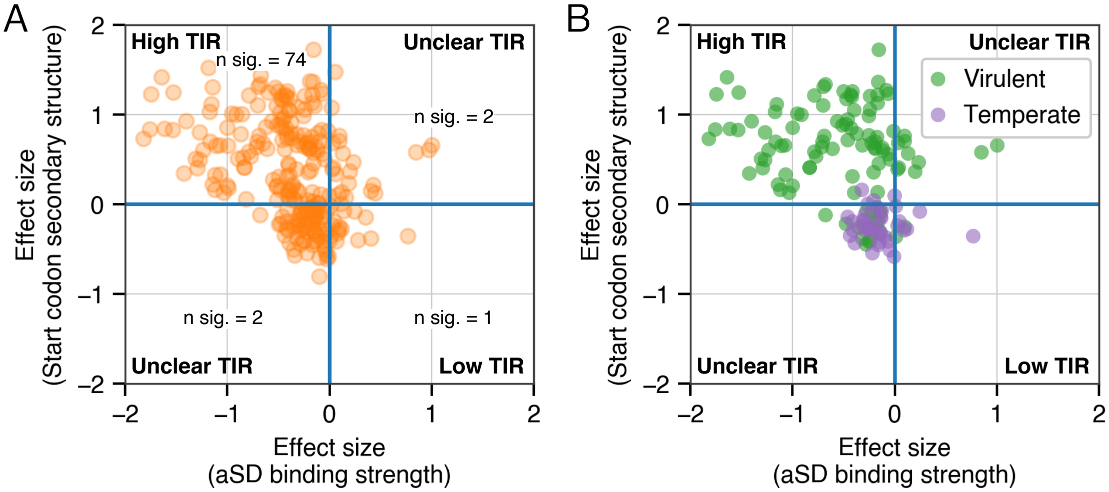
* Corresponding author



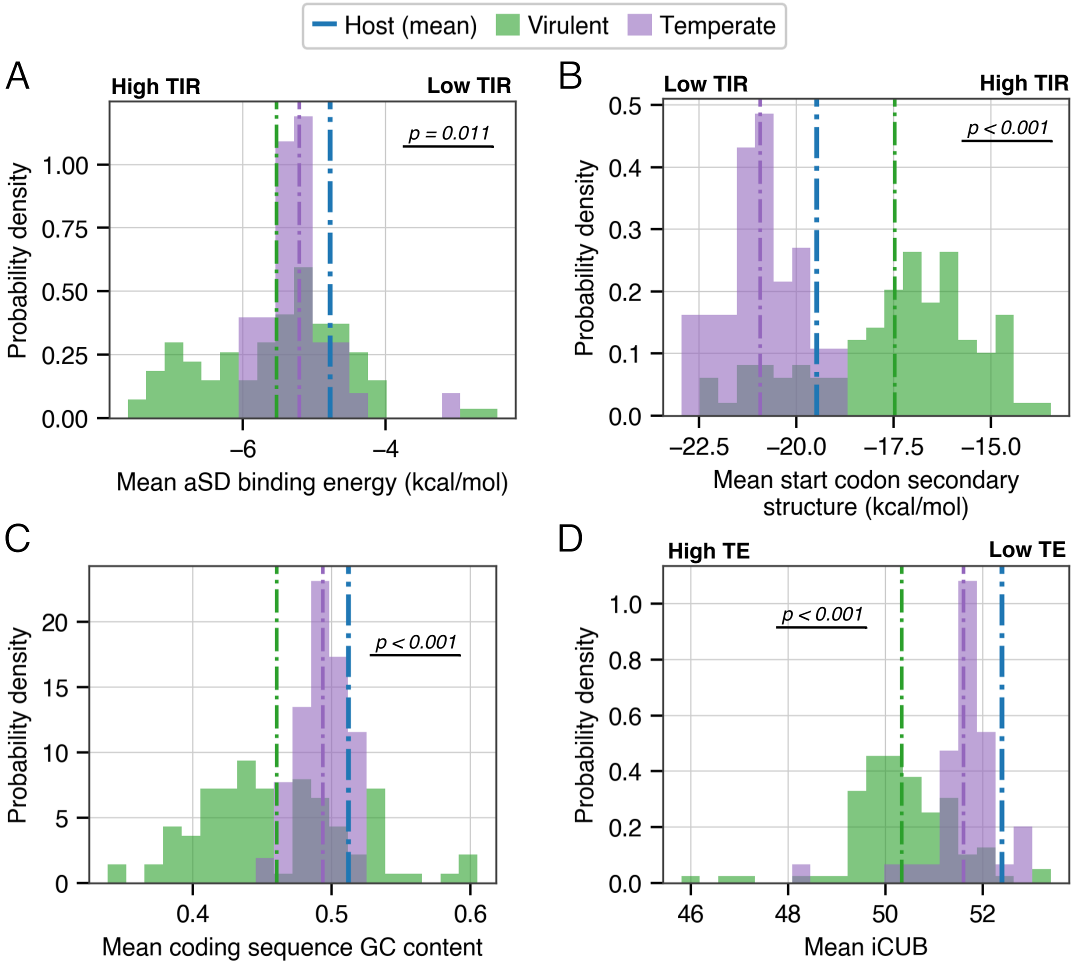
Supplementary Figure S1. Comparing translation-initiation features in phage genomes to host genome subsets. (A) The distribution of mean aSD binding energy across phage genomes compared with indicated subsets of host genes (error bars depict standard errors for host genome subset means, orange square indicates the median phage genome value). The number of phage genomes that are significantly different from gene categories are shown on the right, with red bars extending to the left of 0 indicating genomes with significantly stronger aSD binding strengths and brown bars to the right indicating significantly weaker aSD binding strengths (significance was defined via Welch's t-test with FDR-correction, $p < 0.01$). (B) As in (A), depicting mean start codon secondary structure around the start codon.



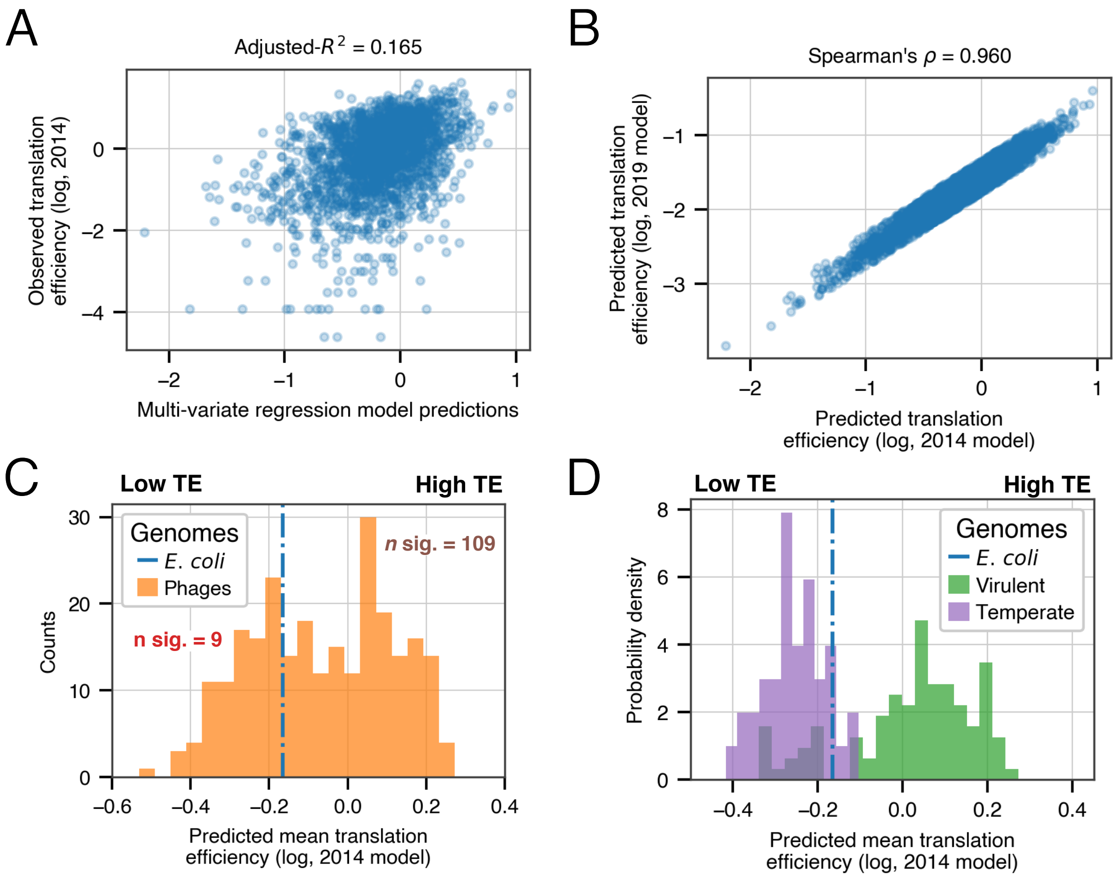
Supplementary Figure S2. GC content and codon usage bias variation in *E. coli* infecting phage genomes. (A) GC content of all phage T7 and *E. coli* coding sequences. (B) As in (A), showing codon usage bias variation (measured with iCUB). (C) Distribution of mean GC contents across all phage genomes, with the host genome mean depicted as a dashed blue line and the number of significant comparisons on either side annotated as 'n. sig'. (D) As in (C), showing the mean iCUB values for each phage genome contrasted with the *E. coli*. In panels (B) and (D) the abbreviation TE refers to translation efficiency, indicating that lower iCUB values are associated with stronger translational selection. A similar correspondence GC content and translation efficiency or translational selection is not known.



Supplementary Figure S3. Multi-variable modeling of translation-initiation differences between phage and host genomes without covariates. (A) Model coefficients (standardized conditional log-odds ratio) from logistic regression comparing sequence features in translation-initiation regions between individual phage and *E. coli* genomes. Coefficients for aSD binding energies and mRNA secondary structure in the start codon region are depicted on the x- and y-axes, respectively. (B) As in (A), with phages colored according to their predicted lifestyle for the subset of phages with high-confidence lifestyle predictions (≥ 0.95 probability of correct assignment using BACPHLIP).



Supplementary Figure S4. Single variable distributions for 143 *E. coli*-infecting phages with confident lifestyle predictions. (A) Mean aSD sequence binding energy, with virulent phages shown in green and temperate phages in purple. In both cases the dashed colored lines show distribution means, and the dashed blue line indicates the *E. coli* genome-wide mean. Further subplots show: (B) mean start codon region secondary structure (C) mean coding sequence GC percentage and (D) mean iCUB values (akin to the effective number of codons with lower values representing stronger codon usage biases and increased predicted translation initiation rates. In all cases, indicated *p*-values represent Welch's t-test comparing temperate and virulent phage distributions (with the host mean shown only as a reference).



Supplementary Figure S5. Differences in translation efficiency across phage lifestyles. (A) Scatter plot of predicted vs observed translation efficiency (log-scaled) using translation efficiency data from Li *et al.* (2014) [59] and a multi-variable linear regression model that includes only aSD sequence binding strength and mRNA secondary structure around the start codon. (B) Comparing model predictions across the entire *E. coli* genome when fit to two separate translation efficiency datasets (2014 refers to Li *et al.* (2014) [59] while 2019 refers to Gorochowski *et al.* (2019) [60]. (C) Differences in mean predicted translation efficiency (using the 2014 model) between phage genomes and the host *E. coli* genome. 'n. sig' denotes the number of phage genomes that are significantly differenent from *E. coli* in either direction (Welch's t-test with FDR-correction, $p < 0.01$). (D) As in panel (C), splitting phage genomes according to lifestyle (including only genomes with high-confidence lifestyle predictions: 39 temperate and 104 virulent). Among virulent phage genomes, 73 (0) have significantly higher (and lower) predicted translation efficiency compared to *E. coli*. By contrast, 0 (3) temperate phage genomes have significantly higher (and lower) predicted levels of translation efficiency.