



eCOMMONS

Loyola University Chicago  
Loyola eCommons

Bioinformatics Faculty Publications

Faculty Publications

5-29-2018

## The Evolution of Molecular Compatibility between Bacteriophage $\Phi$ X174 and its Host

Alexander Kula  
*University of Miami*


Joseph Saelens  
*Duke University*

Alyxandria M. Schubert  
*Center for Biologics Evaluation and Research, U.S. Food and Drug Administration*

Michael Travisano  
*University of Minnesota*

Catherine Putonti  
*Loyola University Chicago, cputonti@luc.edu*

Follow this and additional works at: [https://ecommons.luc.edu/bioinformatics\\_facpub](https://ecommons.luc.edu/bioinformatics_facpub)

 Part of the [Bioinformatics Commons](#), [Biology Commons](#), and the [Virology Commons](#)

### Recommended Citation

Kula, Alexander; Saelens, Joseph; Schubert, Alyxandria M.; Travisano, Michael; and Putonti, Catherine. The Evolution of Molecular Compatibility between Bacteriophage  $\Phi$ X174 and its Host. *Scientific Reports*, 8, : , 2018. Retrieved from Loyola eCommons, Bioinformatics Faculty Publications, <http://dx.doi.org/10.1038/s41598-018-25914-7>

This Article is brought to you for free and open access by the Faculty Publications at Loyola eCommons. It has been accepted for inclusion in Bioinformatics Faculty Publications by an authorized administrator of Loyola eCommons. For more information, please contact [ecommons@luc.edu](mailto:ecommons@luc.edu).



This work is licensed under a [Creative Commons Attribution 4.0 License](#).

© The Authors 2018

# SCIENTIFIC REPORTS



OPEN

## The Evolution of Molecular Compatibility between Bacteriophage $\Phi$ X174 and its Host

Alexander Kula<sup>1,6</sup>, Joseph Saelens<sup>2,7</sup>, Jennifer Cox<sup>1</sup>, Alyxandria M. Schubert<sup>2,8</sup>, Michael Travisano<sup>3,4</sup>  & Catherine Putonti<sup>1,2,5</sup> 

Viruses rely upon their hosts for biosynthesis of viral RNA, DNA and protein. This dependency frequently engenders strong selection for virus genome compatibility with potential hosts, appropriate gene regulation and expression necessary for a successful infection. While bioinformatic studies have shown strong correlations between codon usage in viral and host genomes, the selective factors by which this compatibility evolves remain a matter of conjecture. Engineered to include codons with a lesser usage and/or tRNA abundance within the host, three different attenuated strains of the bacterial virus  $\Phi$ X174 were created and propagated via serial transfers. Molecular sequence data indicate that biosynthetic compatibility was recovered rapidly. Extensive computational simulations were performed to assess the role of mutational biases as well as selection for translational efficiency in the engineered phage. Using bacteriophage as a model system, we can begin to unravel the evolutionary processes shaping codon compatibility between viruses and their host.

The spread of emergent viral diseases critically depends upon rapid adaptation to novel hosts<sup>1</sup>. Analogous to that observed in natural populations (e.g., ref.<sup>2</sup>), experimental evolution of viral pathogens has also demonstrated rapid adaptation to new hosts<sup>3</sup>. Nevertheless, the molecular basis of host specificity within viruses remains contentious<sup>4–8</sup>. While selection experiments have shown that single nucleotide changes can be sufficient to facilitate a viral host shift (e.g., ref.<sup>9</sup>), bioinformatic surveys repeatedly show a high degree of genomic correspondence between viral pathogens and their hosts<sup>4,10</sup>. This is most evident within bacteriophage (phage) species<sup>7,11</sup>. For instance, codon usage of coliphages generally reflects the biased usage of their host which itself reflects the most abundant cognate tRNAs available within host cells<sup>12–14</sup> and mRNA levels<sup>15</sup>. This correspondence of phage and host codon usage is not surprising given that viruses are frequently, often entirely, reliant on their hosts for biosynthesis. This dependency engenders strong selection for virus genome compatibility with potential hosts, a necessity for a successful infection.

While once referred to as “silent”, we now know that synonymous mutations can have a profound effect on both an organism’s phenotype and fitness<sup>16–18</sup>. Deviations from neutral expectations of codon usage can be the result of selection for translational efficiency and/or accuracy, mutational biases, drift, control of gene expression, and structure<sup>19–28</sup>. Reduced viral fitness has been detected in molecular engineering of viral codons via synonymous mutations<sup>29–41</sup> (also see reviews<sup>42,43</sup>). These fitness losses are largely attributed to a reduction of genome translation and show that codon engineering is a promising avenue for generating new vaccines<sup>29–33,36–45</sup>. While the immediate cost of synonymous mutations on viral fitness has been observed, causes and consequences of sequence-specific host adaptation remains elusive.

Phages provide an ideal model system for exploring the evolution of codon usage bias. The literature is rich with experimental evolution of phages, applying various forms of selection<sup>44–50</sup>. Furthermore, substantial bioinformatic analysis of codon usage within phage<sup>7,11</sup> and bacterial<sup>19</sup> genomes has been conducted. Through

<sup>1</sup>Department of Biology, Loyola University Chicago, Chicago, IL, USA. <sup>2</sup>Bioinformatics Program, Loyola University Chicago, Chicago, IL, USA. <sup>3</sup>Department of Ecology, Evolution and Behavior, University of Minnesota, Saint Paul, MN, USA. <sup>4</sup>BioTechnology Institute, University of Minnesota, Saint Paul, MN, USA. <sup>5</sup>Department of Computer Science, Loyola University Chicago, Chicago, IL, USA. <sup>6</sup>Present address: Department of Biology, University of Miami, Coral Gables, FL, USA. <sup>7</sup>Present address: Department of Molecular Genetics and Microbiology, Duke University, Durham, NC, USA. <sup>8</sup>Present address: Division of Bacterial, Parasitic, and Allergenic Products, Office of Vaccines Research and Review, Center for Biologics Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, USA. Correspondence and requests for materials should be addressed to C.P. (email: [cputonti@luc.edu](mailto:cputonti@luc.edu))

Strain Name	Gene Targeted	Engineering	% Deoptimized
S	F	11 synonymous mutations	29.3
E	F	11 synonymous mutations; 1 nonsynonymous mutation	32.9
J	J	12 synonymous mutations	29.0

**Table 1.** Summary of strains created through codon engineering.

experimental evolution of the codon-based attenuated T7 phage, fitness recovery was observed by evolutionary changes in codon use<sup>34</sup>. More rapid rescue has been observed in the passage of codon deoptimized eukaryote-infecting viruses having smaller RNA-based genomes<sup>35,36</sup>. The effects of codon deoptimization on fitness and the recovery of fitness, however, varies from one virus to the next<sup>51</sup>. As prior evidence has shown, synonymous mutations specifically introduced in species having small, compact genomes can have a profound impact on a species' fitness<sup>52–54</sup>. The mechanisms that lead to pathogen-host genome compatibility remain uncertain, leaving the causal factors open questions.

We performed long-term experimental evolution to determine how and at what rate virus-host codon usage evolves, using engineered phage genomes. The coding sequence of the bacteriophage  $\Phi$ X174 was targeted, replacing wild-type codons with deoptimized (relative to its host, *Escherichia coli* C) codons. Three different engineered strains, targeting two different coding regions within the  $\Phi$ X174 genome, were created. The  $\Phi$ X174 genome is small and compact, encoding for just 11 genes in the 5386 nucleotide ssDNA, circular genome; furthermore,  $\Phi$ X174 is known to be sensitive to mutations<sup>54,55</sup>. The combination of engineered sequence changes allows for the simultaneous examination of the role of selection to affect sequence specific adaptation, specifically rates of reversion and translational efficiency within the *E. coli* host. Complementing our experimental efforts, extensive computational simulations were performed to assess the role of selection for translational efficiency. This multi-disciplinary approach provides insight into how genome compatibility arises.

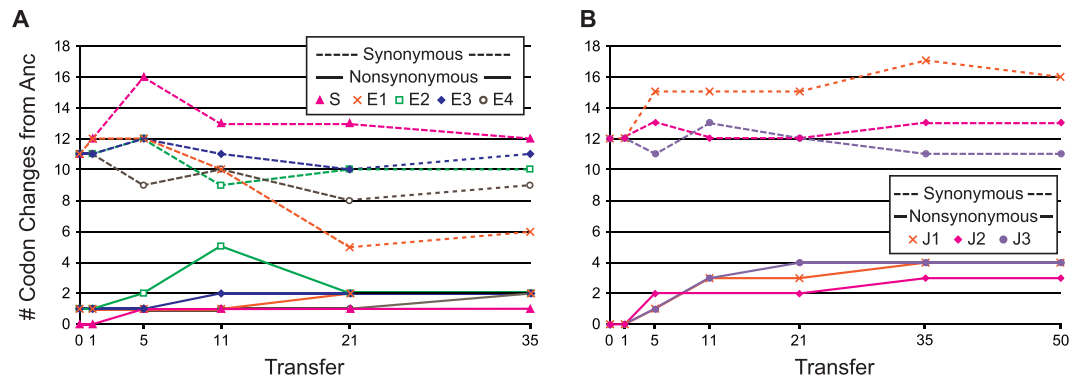
## Results

**Conservation of host genomic compatibility within microviruses.** Codon usage within homologous gene sequences of  $\Phi$ X174 and its two known closest relatives (G4 and  $\alpha$ 3) was examined. Despite only modest sequence similarity, orthologs are similar in their usage of codons favored within the highly expressed genes (HEGs) of their host, *E. coli* (Supplementary Fig. S1). This is true not only of the RefSeq sequences, but also microviruses isolated from environmental samples (results not shown). Furthermore, this trend was also observed within the more distant relative of  $\Phi$ X174:  $\Phi$ MH2K.  $\Phi$ MH2K, also a microvirus, infects *Bdellovibrio bacteriovorus*. The variance of the estimated translation rate between genes was statistically significant ( $p$ -value = 0.00009) while the variance between the species was not. Thus, the observed level of gene-host codon compatibility in these microviruses is conserved regardless of the host species. The homologous coding regions for the F and J coding regions include a codon usage most congruent to their respective host's codon usage biases (Supplementary Fig. S1) and thus were selected for subsequent experimental examination.

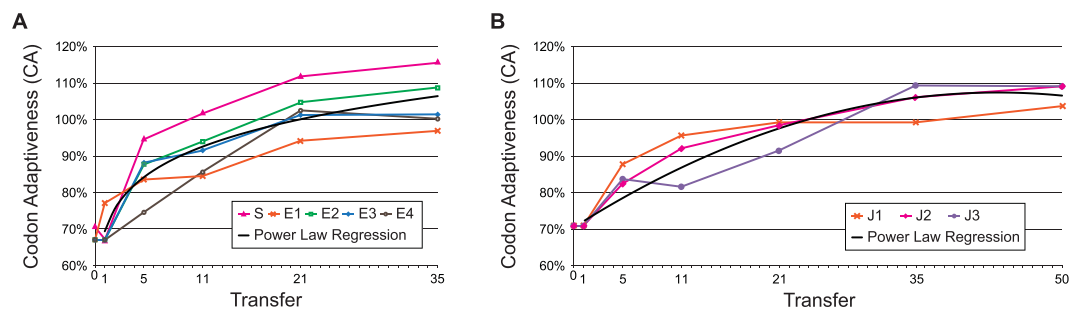
**Strain engineering.** A 66 bp region within the  $\Phi$ X174 capsid protein F coding region and a 69 bp region within the core protein J coding region were re-engineered to include alternate codons, often codons less preferred by the host, such that both regions were comparably deoptimized relative to the ancestral strain (see Methods; summarized in Table 1). Engineered mutants were created from a  $\Phi$ X174 strain in our lab which was well-adapted to the growth conditions employed in the selection experiments carried out here. Two engineered mutants, S and E, were created for the F protein coding region. The S strain contains eleven synonymous substitutions within the 22 codon region (Supplementary Table S1); nine were achieved by single third position changes, while the remaining two codon substitutions included two base changes (first and second position Leucine). The E strain contained these same eleven synonymous substitutions in addition to one nonsynonymous codon replacement (Supplementary Table S1); this particular codon was chosen as sequenced  $\Phi$ X174 strains vary in the amino acid encoded (histidine or arginine). Similarly, a deoptimized sequence was designed for a 23 codon region in J, henceforth referred to as the J strain. The J strain contains twelve synonymous substitutions; all substitutions are achieved by single third position changes (Supplementary Table S2).

The S and E strains were propagated for 35 transfers. While a single propagation of the S strain was performed, the E strain was propagated in quadruplicate. The J strains were propagated for 50 transfers, in triplicate. Four replicates of the Anc strain, the unaltered ancestral strain, were also propagated serving as a control. (Further details regarding the experimental design are included within the Methods.) To distinguish between the engineered genomic sequences and the evolved genomic sequences, the following notation will be used. The engineered mutant strains prior to propagation are referred to as the “S strain”, “E strain”, and “J strain” created from the ancestral “Anc strain”. The serially passaged S, E and J strains are denoted as the S, E, and J lines, collectively referred to as the engineered lines, with replicates denoted by number. The propagated Anc strain is henceforth referred to as the C1, C2, C3, and C4 lines. As anticipated, initial plating of the engineered strains created here showed a significant reduction in the number of successful infections relative to the propagated Anc strain, as measured by plaque forming units (PFU) and burst size (Supplementary Fig. S2).

**Responses to selection.** The targeted region was sequenced for all engineered lines from isolates collected after the 1<sup>st</sup>, 5<sup>th</sup>, 11<sup>th</sup>, 21<sup>st</sup>, and 35<sup>th</sup> transfer; additionally, the J lines were sequenced after the 50<sup>th</sup> transfer. Synonymous, as well as nonsynonymous, mutations occurred both for the codons that were initially manipulated



**Figure 1.** Number of synonymous and nonsynonymous codon differences between the engineered lines and the Anc strain over the course of the selection experiment: **(A)** F coding region engineered lines S and E1, E2, E3, and E4. **(B)** J coding region engineered lines J1, J2, and J3.

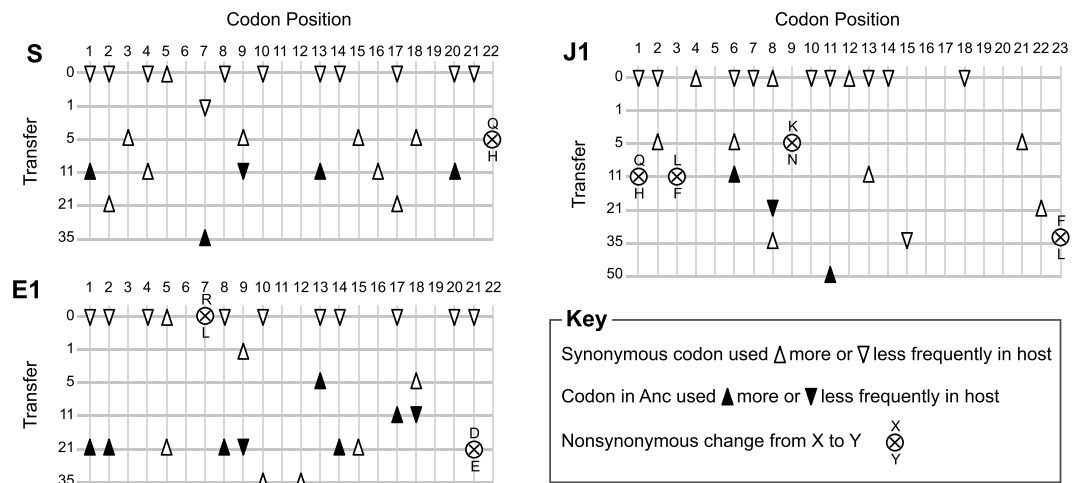


**Figure 2.** Codon adaptiveness (CA) of each of the engineered lines: **(A)** F coding region engineered lines S and E1, E2, E3, and E4. **(B)** J coding region engineered lines J1, J2, and J3.

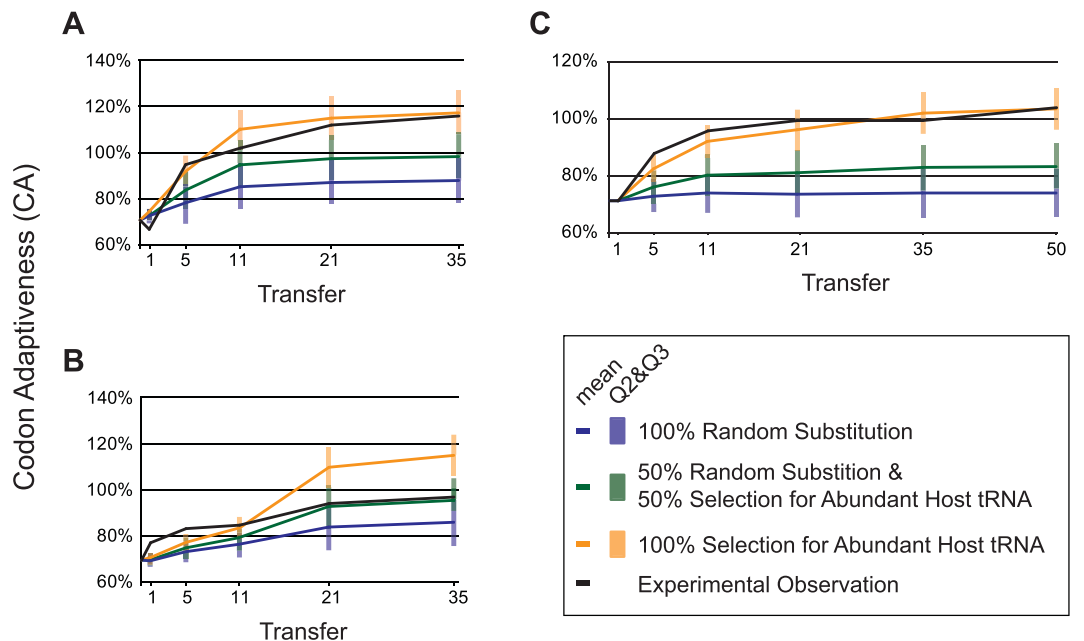
as well as other codons in the region targeted in the engineered lines (Fig. 1). While the E2 line collected after the 11<sup>th</sup> transfer shows the most nonsynonymous differences, five, by the next sampling many of these differences were no longer present in the population. Although several of the codons fixed within the engineered lines were those that were present within the Anc strain, this was not a general result. For each of the mutations identified, changes in codons were identified and assessed relative to the codon usage within the HEGs of *E. coli* C. For all engineered lines, the majority of the mutations result in a substitution for a codon more frequently used within *E. coli* C's set of HEGs (shown in Supplementary Tables S1 and S2).

To assess the putative effects of these mutations on the protein's translational efficiency, we examined the codon adaptiveness (CA) of the targeted windows for each engineered line. This metric represents the individual engineered line's usage of host-preferred codons relative to this same window in the Anc strain. Both the region within the F engineered lines as well as the region within the J engineered lines showed a consistent increase in CA over the course of selection (Fig. 2). The exception being the acquisition of a single synonymous mutation in the S line after 1 transfer and the J3 line after 11 transfers; the CA value of these lines, however, was rapidly improved by the next sampling. At the end of the selection experiment, seven of the eight engineered lines include codons which are utilized more frequently within the host's set of HEGs than are present within this same window in the Anc strain (CA > 100%). Only the F coding region engineered line E1 did not exceed 100%. As the extension of the J lines for an additional 15 transfers reveals, the rate of change in the CA value diminishes (Fig. 2B). In parallel to the steady rise in the CA value, all of the engineered lines showed fitness improvements, with respect to both plaque formation and burst size (Supplementary Fig. S2). Figure 3 illustrates the individual mutations for three of the evolved lines; the remaining lines are shown in Supplementary Fig. S3 and a full listing of the mutations can be found in Supplementary Tables S1 (for the S and E lines) and S2 (for the J lines).

The simultaneous propagation of the C lines provides insight into the probability of mutations arising within the targeted regions of the engineered lines as a result of the selection experiment. The F and J protein coding regions were also sequenced for the C line after the 1<sup>st</sup>, 5<sup>th</sup>, 11<sup>th</sup>, 21<sup>st</sup>, 35<sup>th</sup>, and 50<sup>th</sup> transfers. No nonsynonymous mutations were detected within the J protein coding region. One was observed within the F protein coding region, at genome position 1727 (L242F). This nonsynonymous mutation was first detected after the 21<sup>st</sup> transfer and became fixed in the population; a synonymous mutation at this same position was detected as early as the 5<sup>th</sup> transfer. This nonsynonymous mutation, however, is not unique to the evolved line; of 67 publicly available genomes in GenBank (Supplementary Table S3), 15 have Leucine (including the Anc strain) while the remaining 52 have Phenylalanine at this position.



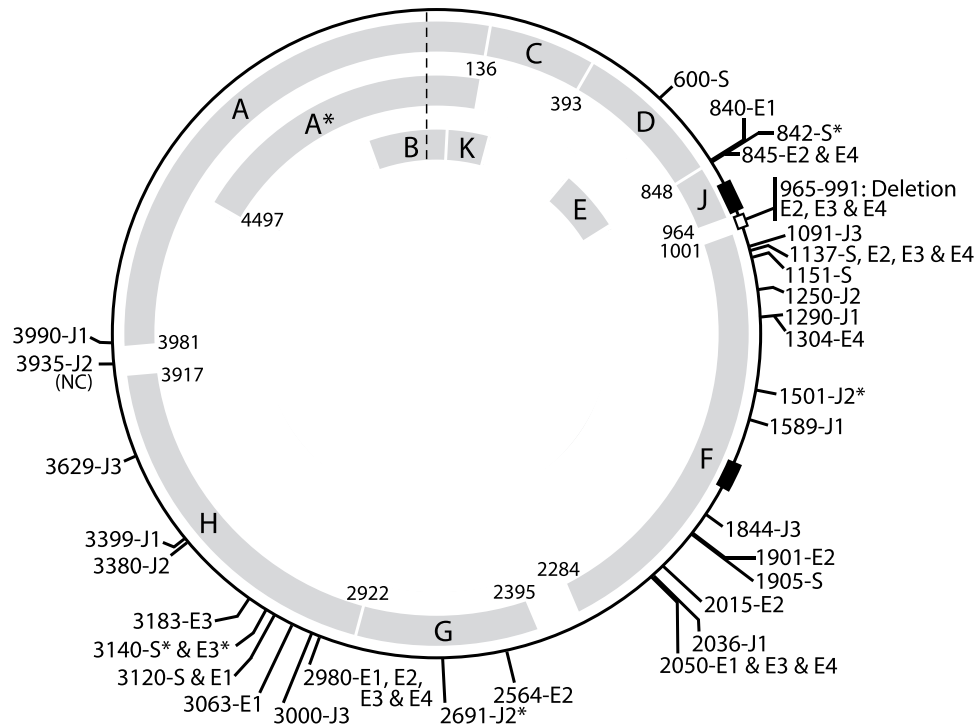
**Figure 3.** Codon changes observed within the engineered lines (relative to the codon in the Anc strain) over the course of the selection experiment. Synonymous mutations are indicated by triangles; those resulting in a codon more frequently used in the *E. coli* C HEGs are indicated by a “ $\Delta$ ” and the converse by a “ $\nabla$ ”. If the mutation results in the codon present within the Anc strain, the triangle is solid black. Nonsynonymous mutations are indicated by  $\otimes$ .



**Figure 4.** Average CA values predicted over time from simulations under three different variations of the role of translational selection and random mutation. The CA values from the experimental assays are shown by the black line. Each of the engineered lines is shown separately, (A) S, (B) E1, and (C) J1.

**Unraveling the selective forces increasing the lines’ CA.** We investigated mutational effects using simulation. The simulation included the effects of random mutation and selection for translational efficiency (see Methods). Conducting 1000 replicates captured the landscape of mutations which could be explored by each engineered sequence. Comparison of the simulations and the experimental assays are shown for the S, E1, and J1 engineered lines in Fig. 4. (The remaining lines can be found in Supplementary Fig. S4.) As the number of mutations increases, increasing divergence is observed in the average CA values for a sequence under strong selection for translational efficiency (the 100% Selection for More Abundant Host tRNA model in yellow) and in its absence (the 100% Random Substitution model in blue). Even when mutations are introduced randomly, the CA value increases because the engineered sequences were severely deoptimized; however, in no case was the random model sufficient to recover the observed increases in codon adaptiveness.

The simulations for the S and E lines suggest that selection for translational efficiency is important in shaping the codon usage of all five of the engineered lines. While the role of selection for translational efficiency



**Figure 5.** Compensatory mutations observed in all engineered lines. Asterisks indicate a synonymous mutation was observed; all other changes in a residue resulted in a nonsynonymous mutation. The open block indicates the location of the excision which occurred in the E2, E3 and E4 lines. The black blocks indicate the location of the 66 bp and 69 bp regions targeted within the S and E and J engineered strains, respectively. “NC” indicates a mutation within the noncoding region of the genome.

between the selected lines may vary, it is not sufficient to explain the number of mutations, reversals, or dN/dS rate. The mutational dynamics in the E1 (Fig. 4B) and E2 and E3 (Supplementary Fig. S4) lines indicate that translational selection is unlikely to be the only factor shaping their codon usage, following the mixed model (shown in Fig. 4 in green). In contrast, the experimental results for the three engineered J lines (Fig. 4C and Supplementary Fig. S4) mirror the expectations under strong selection to utilize codons more frequently within the host’s set of HEGs (yellow lines). Thus, biosynthetic compatibility appears to be a significant source of selection for these lines. The S (Fig. 4A) and E4 (Supplementary Fig. S4) lines also suggest that translational selection plays an important role in shaping its codon usage, albeit not as strong as within the engineered J lines.

**Epistasis.** We hypothesized that there would be sequence changes in other coding regions over the course of selection as a result of protein-protein interactions. Complete genome sequencing was performed for the final isolates of all engineered lines as well as intermediate populations for the S and E lines (see Methods). Figure 5 illustrates the mutations identified within the S, E, and J engineered lines over the course of the selection experiment. The majority of the mutations observed occurred within the structural proteins, regardless of the region engineered. Three of the E lines (E2, E3, and E4) collected after the 35<sup>th</sup> transfer include the excision of 27 nucleotides within the noncoding region between the J and F genes. In order to pinpoint when this excision occurred, isolates from the 22<sup>nd</sup> through 34<sup>th</sup> transfers were assayed for this excision via PCR of this region of the genome. The excision arose in the E2 line in the 27<sup>th</sup> transfer, in the E3 line in the 29<sup>th</sup> transfer, and in the E4 line in the 30<sup>th</sup> transfer. A full listing of the mutations observed outside of the engineered regions throughout the course of the selection experiment within the eight engineered lines can be found in Supplementary Table S4.

By comparing the genetic variability of extant ΦX174 populations (Supplementary Table S3) with the experimental populations evolved here revealed shared diversity. Five of the nonsynonymous mutations within the engineered lines (four within the S and E lines and one within the J lines) (Supplementary Table S4) and all of the nonsynonymous mutations within the control lines (C1, C2, C3, and C4) (Supplementary Table S5) have previously been detected. In addition, through sequence analysis we find that many of these sites are highly variable, in particular the mutations observed within the control lines (Supplementary Table S5), such that numerous different codons and amino acids are exploited in viable strains. Furthermore, the 27 nucleotide excision observed within the E2, E3, and E4 lines is not unique; complete genomes in GenBank also contain this deletion. In fact, there is a correspondence between one of these mutations, A16V in the H protein coding region, and the deletion, both in our lines and those available from NCBI. The remaining nonsynonymous mutations within the engineered lines have not been previously detected within the genomes examined.

## Discussion

Through direct molecular manipulation, we investigated codon bias as it evolved. The bacteriophage  $\Phi$ X174 was genetically engineered to have non-optimal codons resulting in low fitness. Over the course of the selection experiment, however, fitness increased in parallel with the incorporation of more host preferred codons (Fig. 1, Supplementary Fig. S2). Both synonymous and nonsynonymous mutations were observed within the targeted regions as the engineering of synonymous substitutions likely permitted the evolving virus to explore alternative paths within sequence space. This theory and the emergence of novel nonsynonymous mutations within the evolved engineered lines is not exclusive to our study; similarly, throughout the passage of codon deoptimized HIV strains novel nonsynonymous mutations were often frequent<sup>35</sup>. By targeting two individual coding regions of  $\Phi$ X174 – the F and the J coding regions – we were able to determine that the response in phage-host codon compatibility observed was not a result of a particular gene or region selected. The engineered lines fixed between two (line J1) to six (lines E1, E3, and E4) reversions to the un-engineered ancestral codon. However, the reversions were almost exclusively divergent across lineages, with an exception of a single reversion, codon position 1, within the lines in which the F coding region was targeted (Fig. 3). Across the evolving lines, the majority of mutations within the targeted regions were for codons more frequently used within the host's HEGs (Fig. 3). These results demonstrate parallel evolution in a molecular trait: phage-host codon compatibility.

Comparing codon usage within the individual targeted windows between the engineered lines and the Anc strain, we observed an increase in codon adaptiveness over the course of the selection experiment (Fig. 2). Much of the recovery of codon adaptiveness occurred early in the selection experiment regardless of the region targeted. Virtually all improvement in codon adaptiveness occurred within 21 transfers across all evolving lineages, and five of the lines (S, E1, E2, E3, and J1) had ~50% improvements within the first 5 transfers. The increase in CA was not restricted to evolution of the engineered codons, but also involved other codons in the targeted region, indicating that selection was not specific to those codons that were initially engineered. The evolved lines presented here provide empirical evidence that attenuation via codon deoptimization is not permanent, congruent with prior assessments of similar studies<sup>51</sup>.

Caution is however necessary in interpreting the evolutionary basis for increases in codon adaptiveness. The engineered sequence was severely codon deoptimized, and many mutations could have resulted in an increased CA value. The simulations performed under a strictly random substitution model (Fig. 4, blue lines) capture this consequence; the average final CA recovery under this model was 10%. Still, the rapid increase in CA over the course of the selection experiment across all lineages suggests that translational efficiency is a contributing factor shaping genome composition over time. The results of our simulations further support this conjecture. The experimental observations, in particular those of the J engineered lines, most closely fit models incorporating significant selection for more abundant host tRNAs (Fig. 4). The simulations uncover not only the landscape of mutations which could be explored by the engineered sequence but also the selective factors by which phage-host codon usage compatibility evolves. Similar to the results observed here, other studies which saw rapid virus-host codon compatibility recovery also observed fitness recovery<sup>35,36</sup>.

The extent to which virus and host sequences are compatible varies between genes (Supplementary Fig. S1), suggesting that it is well-tuned at a genomic level. Just as viral codon deoptimization can reduce viral fitness, so too can optimization of natively 'non-optimal' genes<sup>56</sup>. In fact, genome manipulation alone is known to have fitness effects<sup>57,58</sup>. Nevertheless, the reduced fitness observed for the engineered S, E, and J strains created here and other studies of codon deoptimization<sup>29–40</sup> are unlikely to be solely due to reduced translational efficiency. Codon engineering may lead to protein and mRNA misfolding or effect genome packaging, genome-capsid interactions, and protein-protein interactions. Even for the model bacteriophage  $\Phi$ X174, many of the aforementioned processes are not fully understood. For instance, while it is known that some of the amino acids within the 66 bp targeted region of F interact with J, G, and F protein subunits during capsid formation<sup>59</sup>, only one nonsynonymous mutation (Q254H in the F coding sequence) occurred within a recognized protein-protein interaction site. Nonsynonymous mutations outside of the engineered region were observed in the evolved S, E, and J lines. This observation is not unique to the engineering of  $\Phi$ X174 as other studies have likewise detected mutations outside of codon-modified segments<sup>34–36</sup>. The 12 nonsynonymous mutations and 10 nonsynonymous mutations in the S and E lines and the J lines, respectively, have not previously been observed in  $\Phi$ X174 genomes. As these nonsynonymous mutations primarily occurred in structural proteins, they may have arisen in response to conformational changes in the engineered regions due to the initial molecular engineering and subsequent evolutionary response.

Isolating the contributions of selection for translational efficiency from those of translational accuracy, mutational bias, and drift has been the subject of decades of intense research activity<sup>13,20–22,60–64</sup>. Exploration of different phage-host systems provides a greater understanding into the evolution of codon usage within viruses. Using the  $\Phi$ X174 system, we investigated translational efficiency in a small virus that does not encode its own tRNAs, as some larger phages do<sup>65</sup>, and is thus entirely dependent upon its host for biosynthesis. We observed rapid evolutionary responses that involved large increases in codon adaptiveness and fitness. While prior studies evolving codon-engineered phages have not observed such a rapid recovery<sup>34</sup>, we hypothesize that the rate of response is influenced by the genome itself – its size, topology, and composition. For instance, the physical constraints of single stranded genomes<sup>66</sup> may contribute to the difference observed between the slow response observed in codon-modified T7 lines (dsDNA genomes) and the  $\Phi$ X174 lines. The long-term evolution of the three engineered  $\Phi$ X174 lines presented here provides the first empirical evidence of rapid selection for genome compatibility in a phage.

Exploring selection for virus-host genome compatibility in phages has two immediate benefits: it provides a model for engineering in viruses infective of eukaryotic cells (vaccine development) and engineering of phages for therapeutic use (phage therapy). The consistent increase in codons frequently utilized in the highly expressed genes of its host *E. coli* C suggests that translational efficiency was important during selection. The response

observed was in all engineered strains and amongst all replicate lines. Thus, the consistent increase in virus–host genomic compatibility observed is a genome phenomenon rather than a residual of engineering within a specific region/gene. This is further supported through the computational simulations performed. The results provide insight into the tempo and mode in which viruses adapt in response to available hosts.

## Materials and Methods

**Calculation of codon usage.** The complete sequence and annotation for the *E. coli* C genome (GenBank: NC\_010468) was downloaded from NCBI. The codon frequencies were calculated for the 40 highly expressed gene sequences (HEGs)<sup>19</sup>. The unscaled proportion of codons in each codon family was calculated. In contrast to the relative synonymous codon usage<sup>67</sup>, this value, which we refer to as NRSCU<sup>68</sup>, weights each amino acid equally. NRSCU values were retrieved from the Codon Bias Database<sup>68</sup>.

The genome and annotation files for the viral species  $\Phi$ X174 (GenBank: NC\_001422), G4 (GenBank: NC\_001420),  $\alpha$ 3 (GenBank: NC\_001330), and  $\Phi$ MH2K (GenBank: NC\_002643) were downloaded from NCBI. Comparisons for the  $\Phi$ MH2K–host codon compatibility also required the files, again retrieved from NCBI, for its host *Bdellovibrio bacteriovorus*; the reference genome for the strain HD100 was used (GenBank: NC\_005363). Similarly, the codon usage of the HEGs within the *B. bacteriovorus* genome was calculated. Fig. S1 illustrates the phage–host codon usage compatibility (NRSCU value) for the six homologous coding regions of  $\Phi$ X174, G5,  $\alpha$ 3, and  $\Phi$ MH2K (panel A) and for all 11 homologous genes of  $\Phi$ X174, G5, and  $\alpha$ 3 (panel B).

**Sequence design.** Using the genome sequence for the  $\Phi$ X174 Anc strain (GenBank: AF176034)<sup>45</sup>, the restriction enzyme cut sites within the F capsid coding region were identified with the NEB Cutter online tool<sup>69</sup>; PshAI and AhdI were selected because each recognized unique cut sites within the phage’s genome (at nucleotide positions 1694 and 1765, respectively). A 66 bp (nucleotide positions 1700–1765) region between these two cut sites, 22 codons, was then assessed for the individual codon usage within the *E. coli* C host species according to the codon bias of the HEGs from our calculations (Supplementary Table S1). Two sequences were designed, each containing eleven synonymous substitutions. The S strain includes only these eleven synonymous mutations while the E strain includes the synonymous mutations as well as a single nonsynonymous mutation at genome position 1718–1720. The nonsynonymous mutation of CGC (Arginine) for the least favored codon of Leucine was chosen as it is one of the least conserved residues within the region, as denoted by PDBsum’s residue conservation calculations<sup>70,71</sup>. The J strain targeted the region within the  $\Phi$ X174 Anc strain, position 893–961. The restriction enzyme cut sites within the J coding region were also identified with the NEB Cutter online tool<sup>69</sup>; BstAPI and Sau961 were selected because each recognized cut sites flanking the coding region (at nucleotide positions 898 and 978, respectively). Just as had been performed for the design of the S and E strains, each codon in the J region targeted was compared to the NRSCU value for the *E. coli* genome (Supplementary Table S2). Only synonymous mutations were incorporated within the design of the engineered J strain. The oligos for the engineered sequences were synthesized by and obtained from Eurofins MWG Operon.

**Creation of engineered strain.** The Anc strain was originally obtained from C. Burch (University of North Carolina, NC). This ancestral strain was plated from our freezer stock collection. One plate was harvested for the C line (control) and production of the engineered strains. Genomic extraction was performed using the UltraClean<sup>TM</sup> Microbial DNA Isolation Kit following the standard protocol with an additional heating of the prep for 10 minutes at 70 °C to increase lysis efficiency (as suggested by protocol). Double digests using the corresponding enzymes for the F and J coding regions were conducted following the manufacturer’s protocol (New England Biolabs). The digested DNA was separated by gel electrophoresis through a 1.2% agarose gel. DNA fragments were excised from the gel and purified using the UltraClean<sup>TM</sup> 15 DNA Purification kit. Ligation was performed with 7  $\mu$ l of the digested DNA, 1  $\mu$ l of the synthesized oligo, 1  $\mu$ l ligase 10  $\times$  buffer, and 1  $\mu$ l T4 DNA ligase overnight at 4 °C.

The ligation product (5  $\mu$ l) was incubated with 400  $\mu$ l of *E. coli* C spheroplast for 20 minutes at 37 °C; PAM medium (3 ml, pre-warmed to 37 °C) was added and the preparation was incubated for 90 min. The phage was released using a 1:10 dilution into water then titered. This process was carried out for each strain. Each phage strain’s lysate was then plated as follows: 100  $\mu$ l of phage was added to 3 ml 0.5% agar LB and 1 ml of turbid *E. coli* C culture and then overlaid on a 1.7% agar LB plate. Plates were incubated overnight at 37 °C. Plates were harvested and suspended in 0.8% saline solution and treated with 50  $\mu$ l chloroform. Single plaques were selected for each strain as the initial genotype for the subsequent lines. The genomes of the three engineered strains and the ancestral strain were confirmed by capillary sequencing.

**Propagation of engineered lines.** The host *E. coli* C strain was also obtained from C. Burch (University of North Carolina, NC). Propagations were carried out as follows. One line of the S strain, four replicate lines of the E strain, three replicate lines of the J strain, and four lines of the ancestral strain (Anc) to serve as a control were propagated. While the S and E lines were propagated for 35 transfers, the J and Anc control lines were propagated for an additional 15 transfers. Co-cultures were carried out for seven hours per transfer. The emergence of bacterial resistance was also measured for co-culture of this duration determining that phage-sensitive *E. coli* dominated the population.

Initially LB was inoculated with the host *E. coli* C strain, taken from our frozen stock collection. 2 ml of turbid *E. coli* C cultures in exponential growth was aliquoted into a 13 mm culture tube along with 500  $\mu$ l of phage solution titered such that the initial MOI < 0.001. (Under the conditions described hereafter, bacterial growth curves for our *E. coli* C strain were conducted – quantified both by spectrophotometry and colony counts – to ascertain phases of growth and CFU/mL throughout, results not shown.) The tube was then capped and placed in a shaking incubator at 37 °C for 7 hours after which the tube was treated with 200  $\mu$ l of chloroform and gently vortexed for



5 seconds. Next, 500  $\mu\text{l}$  was collected to inoculate freshly grown *E. coli* C in a new culture tube and 500  $\mu\text{l}$  was collected into a microcentrifuge tube and stored at 4 °C. Every third transfer an additional 100  $\mu\text{l}$  was collected and plated. Phage isolates were plated as described previously; virus lysates were stored both at –80 °C in 50/50 glycerol/water (v/v) as well as at 4 °C.

In an effort to maintain a static *E. coli* C population and thus minimize bacterial resistant to the phage from one transfer to the next, fresh *E. coli* C cultures were made daily from naïve cultures. Prior to inoculation with phage lysate, the naïve *E. coli* C culture was grown to the same density as the initial inoculations.

**Sequencing.** Genomic extraction was performed of a single genotype per collection time using the UltraClean™ Microbial DNA Isolation kit as described previously. Twelve primer pairs were designed using the Primer3 web-application<sup>72</sup>; when all twelve pairs are used, a minimum 2 × coverage of the genome is possible (primer sequences available upon request). PCR products were purified using ExoSAP-It and sequenced by the University of Chicago Cancer Research Center DNA Sequencing Facility.

Sequencing of the complete genome with a 4 × coverage was conducted after the 1<sup>st</sup>, 5<sup>th</sup>, 11<sup>th</sup>, 21<sup>st</sup>, and 35<sup>th</sup> transfers for the S and E lines and after the 50<sup>th</sup> transfer for the J lines. The C1 line was also sequenced after the 1<sup>st</sup>, 5<sup>th</sup>, 11<sup>th</sup>, 21<sup>st</sup> and 35<sup>th</sup> transfers. Sequences of the final evolved lines for the C1 line (GenBank: HM775306), S line (GenBank: HM775307), E1 line (GenBank: HM775308), E2 line (GenBank: HM775309), E3 line (GenBank: HM775310), and E4 line (GenBank: HM775311). The three J lines have been deposited as well (GenBank numbers being processed).

Sequencing at each collection time was conducted initially by extracting viral DNA from lysate. As such, the potential for numerous genotypes to be pooled existed. Additionally, we also plated via serial dilutions lysate collected and selected plaques at random for sequencing. In all cases the same genotype was recovered suggesting relatively low heterogeneity within the population.

**Sequence analysis.** The sequences generated in this study were assembled using LaserGene SeqMan (DNASTAR, Inc.). Comparisons between the isolate contigs to the ancestral strain's sequence were conducted by performing multiple sequence alignments using ClustalW within BioEdit (<http://www.mbio.ncsu.edu/BioEdit/bioedit.html>). Comparisons with environmental samples (Table S3; GenBank: AY751298, DQ079870–2, DQ079874–9907, DQ079909, NC\_007817, NC\_007821, and NC\_007856)<sup>73</sup> were also downloaded from NCBI.

**Codon adaptiveness.** We used the unscaled proportion of codons in each codon family (NRSCU). This metric captures only increases/decreases in the use of host-preferred codons. These values are available through the CBDB site<sup>68</sup>. For each sequenced isolate, the codon adaptiveness or CA value was quantified. This metric represents the individual engineered line's codon usage in comparison to this same window in the Anc strain relative to the codon usage within the host's HEGs. This metric was implemented rather than the codon adaptation index or CAI value<sup>74</sup> which takes length of the sequences into consideration; as we were comparing a region of the same length, length was not a contributing factor.

**Adsorption assays and burst assays.** Plaque forming unit (PFU) counts were conducted by first titrating the viral lysate (via dilution series conducted in triplicate) such that equivalent initial viral concentrations were plated: 100  $\mu\text{l}$  of phage was added to 3 ml 0.5% agar LB and 1 ml of turbid *E. coli* C culture and then overlaid on a 1.7% agar LB plate. Each strain was plated with three replicates and plaques were counted. Adsorption assays were also performed, in triplicate per strain/line. The assay estimates fitness based on the doublings of phage concentration per hour which is not scaled to generation time which may differ among the engineered lines. This allows for a comparison between the Anc strain and evolved strains based on their absolute growth rate with their native host *E. coli* C. The assay is an additional measure of fitness and determines which phage can grow the fastest. *E. coli* C was grown for 90 minutes until visible turbidity was observed. 10 mL of *E. coli* C was inoculated with 1 mL of bacteriophage (titered such that  $\text{MOI} < 0.01$ ) and incubated at 37 °C. After 5 minutes, 1 mL of the culture was removed, microcentrifuged, and the phage within the supernatant was plated via a dilution series; this represents the initial concentration of phage ( $N_0$ ). After 60 minutes ( $t$ ), another 1 mL of the culture was removed and the phage in the supernatant was again isolated and titered. This is considered the final concentration of phage ( $N_t$ ). To find the adsorption rate ( $k$ ), the equation  $N_t = N_0 e^{-kCt}$  can be used where  $C$  is the bacterial cell density<sup>75</sup>. The experiments for determining the adsorption time can also be used to determine the burst size, taking just one pre-lysis data point and one post-lysis data point with multiple replicates<sup>76</sup>.

**Simulations.** Each engineered line was evaluated separately. At each of the five (six in the case of the J lines) time points in which sequencing was performed, the same number of experimentally observed mutations – synonymous and nonsynonymous – were introduced. The CA was then calculated for the synthetically “evolved” sequence. Two strategies for mutation were developed. In the case of the first, nucleotides were mutated with equal probability of substitution for each of the four bases. The second strategy only incorporated a mutation if the change in the codon is for a tRNA that is more abundant in the *E. coli* host (as assessed via the NRSCU values of the two codons). Given the fact that such a small region of the genome was being investigated, the particular nucleotides targeted were selected at random (using Marsaglia's CMWC strategy). Simulations were executed with 1000 replicates per time point per line accounting for varying influences (from 0–100%) of each of the two strategies. Simulations were performed using code developed here in C++ (available upon request).

**Data availability.** Sequence data generated during the current study are available in the GenBank, accession numbers HM775306–HM775311. The three J lines have been deposited as well (GenBank numbers being processed). Sequences analyzed in this study are listed in Supplementary Table S3.

## References

- Antia, R., Regoes, R. R., Koella, J. C. & Bergstrom, C. T. The role of evolution in the emergence of infectious diseases. *Nature* **426**, 658–661 (2003).
- Parrish, C. R. *et al.* Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiol. Mol. Biol. Rev.* **72**, 457–470 (2008).
- Hall, J. P. J., Harrison, E. & Brockhurst, M. A. Viral host-adaptation: insights from evolution experiments with phages. *Curr. Opin. Virol.* **3**, 572–577 (2013).
- Shackelton, L. A., Parrish, C. R. & Holmes, E. C. Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. *J. Mol. Evol.* **62**, 551–563 (2006).
- Pride, D. T., Wassenaar, T. M., Ghose, C. & Blaser, M. J. Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. *BMC Genomics* **7**, 8, <https://doi.org/10.1186/1471-2164-7-8> (2006).
- Greenbaum, B. D., Levine, A. J., Bhanot, G. & Rabadan, R. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. *PLoS Pathog.* **4**, e1000079, <https://doi.org/10.1371/journal.ppat.1000079> (2008).
- Carbone, A. Codon bias is a major factor explaining phage evolution in translationally biased hosts. *J. Mol. Evol.* **66**, 210–223 (2008).
- Novella, I. S., Presloid, J. B., Smith, S. D. & Wilke, C. O. Specific and nonspecific host adaptation during arboviral experimental evolution. *J. Mol. Microbiol. Biotechnol.* **21**, 71–81 (2011).
- Ford, B. E. *et al.* Frequency and fitness consequences of bacteriophage  $\Phi$ 6 host range mutations. *PLoS ONE* **9**, e113078, <https://doi.org/10.1371/journal.pone.0113078> (2014).
- Jenkins, G. M. & Holmes, E. C. The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Res.* **92**, 1–7 (2003).
- Lucks, J. B., Nelson, D. R., Kudla, G. R. & Plotkin, J. B. Genome landscapes and bacteriophage codon usage. *PLoS Comput. Biol.* **4**, e1000001, <https://doi.org/10.1371/journal.pcbi.1000001> (2008).
- Ikemura, T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.* **146**, 1–21 (1981).
- Bulmer, M. Coevolution of codon usage and transfer RNA abundance. *Nature* **325**, 728–730 (1987).
- Dong, H., Nilsson, L. & Kurland, C. G. Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J. Mol. Biol.* **260**, 649–663 (1996).
- dos Reis, M., Wernisch, L. & Savva, R. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. *Nucleic Acids Res.* **31**, 6976–6985 (2003).
- Agashe, D. *et al.* Large-effect beneficial synonymous mutations mediate rapid and parallel adaptation in a bacterium. *Mol. Biol. Evol.* **33**, 1542–1553 (2016).
- Agashe, D., Martinez-Gomez, N. C., Drummond, D. A. & Marx, C. J. Good codons, bad transcript: Large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. *Mol. Biol. Evol.* **30**, 549–560 (2013).
- Bailey, S. F., Hinz, A. & Kassen, R. Adaptive synonymous mutations in an experimentally evolved *Pseudomonas fluorescens* population. *Nat. Commun.* **5**, 4076, <https://doi.org/10.1038/ncomms5076> (2014).
- 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 Res.* **33**, 1141–1153 (2005).
- Hershberg, R. & Petrov, D. A. Selection on codon bias. *Annu. Rev. Genet.* **42**, 287–299 (2008).
- Shah, P. & Gilchrist, M. A. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *Proc. Natl. Acad. Sci. USA* **108**, 10231–10236 (2011).
- Wallace, E. W. J., Airolidi, E. M. & Drummond, D. A. Estimating selection on synonymous codon usage from noisy experimental data. *Mol. Biol. Evol.* **30**, 1438–1453 (2013).
- Cardinale, D. J., DeRosa, K. & Duffy, S. Base composition and translational selection are insufficient to explain codon usage bias in plant viruses. *Viruses* **5**, 162–181 (2013).
- Chithambaram, S., Prabhakaran, R. & Xia, X. Differential codon adaptation between dsDNA and ssDNA phages in *Escherichia coli*. *Mol. Biol. Evol.* **31**, 1606–1617 (2014).
- Chithambaram, S., Prabhakaran, R. & Xia, X. The effect of mutation and selection on codon adaptation in *Escherichia coli* bacteriophage. *Genetics* **197**, 301–315 (2014).
- Ma, M. R. *et al.* Overall codon usage pattern of enterovirus 71. *Genet. Mol. Res.* **13**, 336–343 (2014).
- Shin, Y. C., Bischof, G. F., Lauer, W. A. & Desrosiers, R. C. Importance of codon usage for the temporal regulation of viral gene expression. *Proc. Natl. Acad. Sci. USA* **112**, 14030–14035 (2015).
- Quax, T. E. F., Claessens, N. J., Söll, D. & van der Oost, J. Codon bias as a means to fine-tune gene expression. *Mol. Cell.* **59**, 149–161 (2015).
- Burns, C. C. *et al.* Genetic inactivation of poliovirus infectivity by increasing the frequencies of CpG and UpA dinucleotides within and across synonymous capsid region codons. *J. Virol.* **83**, 9957–9969 (2009).
- Burns, C. C. *et al.* Modulation of poliovirus replicative fitness in HeLa cells by deoptimization of synonymous codon usage in the capsid region. *J. Virol.* **80**, 3259–3272 (2006).
- Mueller, S. *et al.* Live attenuated influenza virus vaccines by computer-aided rational design. *Nat. Biotechnol.* **28**, 723–726 (2010).
- Mueller, S., Papamichail, D., Coleman, J. R., Skiena, S. & Wimmer, E. Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity. *J. Virol.* **80**, 9687–9696 (2006).
- Coleman, J. R. *et al.* Virus attenuation by genome-scale changes in codon pair bias. *Science* **320**, 1784–1787 (2008).
- Bull, J. J., Molineux, I. J. & Wilke, C. O. 2012. Slow fitness recovery in a codon-modified viral genome. *Mol. Biol. Evol.* **29**, 2997–3004 (2012).
- Martrus, G., Nevot, M., Andres, C., Clotet, B. & Martinez, M. A. Changes in codon-pair bias of human immunodeficiency virus type 1 have profound effects on virus replication in cell culture. *Retrovirology* **10**, 78, <https://doi.org/10.1186/1742-4690-10-78> (2013).
- Nougairède, A. *et al.* Random codon re-encoding induces stable reduction of replicative fitness of Chikungunya virus in primate and mosquito cells. *PLoS Pathog.* **9**, e1003172, <https://doi.org/10.1371/journal.ppat.1003172> (2013).
- Le Nouën, C. *et al.* Attenuation of human respiratory syncytial virus by genome-scale codon-pair deoptimization. *Proc. Natl. Acad. Sci. USA* **111**, 13169–13174 (2014).
- Meng, J., Lee, S., Hotard, A. L. & Moore, M. L. Refining the balance of attenuation and immunogenicity of respiratory syncytial virus by targeted codon deoptimization of virulence genes. *MBio* **5**, e01704–01714, <https://doi.org/10.1128/mBio.01704-14> (2014).
- Ni, Y.-Y. *et al.* Computer-aided codon-pairs deoptimization of the major envelope GP5 gene attenuates porcine reproductive and respiratory syndrome virus. *Virology* **450–451**, 132–139 (2014).
- Nogales, A. *et al.* Influenza A virus attenuation by codon deoptimization of the NS gene for vaccine development. *Journal of Virology* **88**, 10525–10540 (2014).
- de Fabritus, L., Nougairède, A., Aubry, F., Gould, E. A. & de Lamballerie, X. Attenuation of tick-borne encephalitis virus using large-scale random codon re-encoding. *PLoS Pathog.* **11**, e1004738, <https://doi.org/10.1371/journal.ppat.1004738> (2015).
- Baker, S. F., Nogales, A. & Martínez-Sobrido, L. Downregulating viral gene expression: codon usage bias manipulation for the generation of novel influenza A virus vaccines. *Future Virol.* **10**, 715–730 (2015).
- Martínez, M. A., Jordan-Paiz, A., Franco, S. & Nevot, M. Synonymous virus genome recoding as a tool to impact viral fitness. *Trends Microbiol.* **24**, 134–147 (2016).

44. Wichman, H. A., Badgett, M. R., Scott, L. A., Boulianne, C. M. & Bull, J. J. Different trajectories of parallel evolution during viral adaptation. *Science*. **285**, 422–424 (1999).
45. Crill, W. D., Wichman, H. A. & Bull, J. J. Evolutionary reversals during viral adaptation to alternating hosts. *Genetics*. **154**, 27–37 (2000).
46. Dennehy, J. J., Friedenber, N. A., Holt, R. D. & Turner, P. E. Viral ecology and the maintenance of novel host use. *Am. Nat.* **167**, 429–439 (2006).
47. Duffy, S., Burch, C. L. & Turner, P. E. Evolution of host specificity drives reproductive isolation among RNA viruses. *Evolution*. **61**, 2614–2622 (2007).
48. Pepin, K. M. & Wichman, H. A. Variable epistatic effects between mutations at host recognition sites in phiX174 bacteriophage. *Evolution*. **61**, 1710–1724 (2007).
49. Meyer, J. R. *et al.* Repeatability and contingency in the evolution of a key innovation in phage lambda. *Science*. **335**, 428–432 (2012).
50. Dover, J. A., Burmeister, A. R., Molineux, I. J. & Parent, K. N. Evolved populations of *Shigella flexneri* phage Sf6 acquire large deletions, altered genomic architecture, and faster life cycles. *Genome Biol. Evol.* **8**, 2827–2840 (2016).
51. Bull, J. J. Evolutionary reversion of live viral vaccines: Can genetic engineering subdue it? *Virus Evol.* **1**, vev005, <https://doi.org/10.1093/ve/vev005> (2015).
52. Novella, I. S., Zárate, S., Metzgar, D. & Ebendick-Corpus, B. E. Positive selection of synonymous mutations in vesicular stomatitis virus. *J. Mol. Biol.* **342**, 1415–1421 (2004).
53. Novella, I. S. *et al.* Genomic evolution of vesicular stomatitis virus strains with differences in adaptability. *J. Virol.* **84**, 4960–4968 (2010).
54. Cuevas, J. M., Domingo-Calap, P. & Sanjuán, R. The fitness effects of synonymous mutations in DNA and RNA viruses. *Mol. Biol. Evol.* **29**, 17–20 (2012).
55. Domingo-Calap, P., Cuevas, J. M. & Sanjuán, R. The fitness effects of random mutations in single-stranded DNA and RNA bacteriophages. *PLoS Genet.* **5**, e1000742, <https://doi.org/10.1371/journal.pgen.1000742> (2009).
56. Villanueva, E., Martí-Solano, M. & Fillat, C. Codon optimization of the adenoviral fiber negatively impacts structural protein expression and viral fitness. *Sci Rep.* **6**, 27546, <https://doi.org/10.1038/srep27546> (2016).
57. Chan, L. Y., Kosuri, S. & Endy, D. Refactoring bacteriophage T7. *Mol. Syst. Biol.* **1**, 2005.0018, <https://doi.org/10.1038/msb4100025> (2005).
58. Springman, R., Molineux, I. J., Duong, C., Bull, R. J. & Bull, J. J. Evolutionary stability of a refactored phage genome. *ACS Synth Biol.* **1**, 425–430 (2012).
59. McKenna, R., Ilag, L. L. & Rossmann, M. G. Analysis of the single-stranded DNA bacteriophage phi X174, refined at a resolution of 3.0 Å. *J. Mol. Biol.* **237**, 517–543 (1994).
60. Stoletzki, N. & Eyre-Walker, A. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Mol. Biol. Evol.* **24**, 374–381 (2007).
61. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science*. **324**, 255–258 (2009).
62. Palidwor, G. A., Perkins, T. J. & Xia, X. A general model of codon bias due to GC mutational bias. *PLoS ONE*. **5**, e13431, <https://doi.org/10.1371/journal.pone.0013431> (2010).
63. Gingold, H. & Pilpel, Y. Determinants of translation efficiency and accuracy. *Mol. Syst. Biol.* **7**, 481, <https://doi.org/10.1038/msb.2011.14> (2011).
64. Zhou, Z. *et al.* Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc. Natl. Acad. Sci. USA* **113**, E6117–E6125 (2016).
65. Bailly-Bechet, M., Vergassola, M. & Rocha, E. Causes for the intriguing presence of tRNAs in phages. *Genome Res.* **17**, 1486–1495 (2014).
66. Tubiana, L., Božič, A. L., Micheletti, C. & Podgornik, R. Synonymous mutations reduce genome compactness in icosahedral ssRNA viruses. *Biophys. J.* **108**, 194–202 (2015).
67. Sharp, P. M., Tuohy, T. M. & Mosurski, K. R. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.* **14**, 5125–5143 (1986).
68. Hilterbrand, A., Saelens, J. & Putonti, C. CBDB: the codon bias database. *BMC Bioinformatics*. **13**, 62, <https://doi.org/10.1186/1471-2105-13-62> (2012).
69. Vincze, T., Posfai, J. & Roberts, R. J. NEBcutter: A program to cleave DNA with restriction enzymes. *Nucleic Acids Res.* **31**, 3688–3691 (2003).
70. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* **38**, W529–533 (2010).
71. de Beer, T. A. P., Berka, K., Thornton, J. M. & Laskowski, R. A. PDBsum additions. *Nucleic Acids Res.* **42**, D292–296 (2014).
72. Untergasser, A. *et al.* Primer3—new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115, <https://doi.org/10.1093/nar/gks596> (2012).
73. Rokyta, D. R., Burch, C. L., Caudle, S. B. & Wichman, H. A. Horizontal gene transfer and the evolution of microvirid coliphage genomes. *J. Bacteriol.* **188**, 1134–1142 (2006).
74. Sharp, P. M. & Li, W. H. The codon adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295 (1987).
75. Bull, J. J., Badgett, M. R., Springman, R. & Molineux, I. J. Genome properties and the limits of adaptation in bacteriophages. *Evolution*. **58**, 692–701 (2004).
76. Hyman, P. & Abedon, S. T. Practical methods for determining phage growth parameters. *Methods Mol. Biol.* **501**, 175–202 (2009).

## Acknowledgements

The authors would like to thank Adam Hilterbrand and Ramunas Stanciasukas for their assistance in fitness assays as well as Dr. Jeffrey Doering for comments regarding the manuscript. CP's work was supported by a Research Support Grant from Loyola University Chicago for the funding of supplies.

## Author Contributions

A.K., J.S., J.C., and A.M.S. conducted the experiments. A.K., J.C., and C.P. conducted the sequence analyses. M.T. and C.P. conceived of the experiment and wrote the main manuscript text. C.P. prepared all figures and tables. All authors reviewed the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-018-25914-7>.

**Competing Interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018