RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies

Citation for published version:
Tulloch, F, Atkinson, NJ, Evans, DJ, Ryan, MD & Simmonds, P 2014, 'RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies' eLIFE, vol 3. DOI: 10.7554/eLife.04531

Digital Object Identifier (DOI):
10.7554/eLife.04531

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
eLIFE

Publisher Rights Statement:
© 2014, Tulloch et al

This article is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use and redistribution provided that the original author and source are credited.

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies

Fiona Tulloch1, Nicky J Atkinson2, David J Evans3, Martin D Ryan1, Peter Simmonds2*

1School of Biology, University of St Andrews, St Andrews, United Kingdom; 2Infection and Immunity Division, Roslin Institute, University of Edinburgh, Edinburgh, United Kingdom; 3School of Life Sciences, University of Warwick, Coventry, United Kingdom

Abstract Mutating RNA virus genomes to alter codon pair (CP) frequencies and reduce translation efficiency has been advocated as a method to generate safe, attenuated virus vaccines. However, selection for disfavoured CPs leads to unintended increases in CpG and UpA dinucleotide frequencies that also attenuate replication. We designed and phenotypically characterised mutants of the picornavirus, echovirus 7, in which these parameters were independently varied to determine which most influenced virus replication. CpG and UpA dinucleotide frequencies primarily influenced virus replication ability while no fitness differences were observed between mutants with different CP usage where dinucleotide frequencies were kept constant. Contrastingly, translation efficiency was unaffected by either CP usage or dinucleotide frequencies. This mechanistic insight is critical for future rational design of live virus vaccines and their safety evaluation; attenuation is mediated through enhanced innate immune responses to viruses with elevated CpG/UpA dinucleotide frequencies rather than the viruses themselves being intrinsically defective.

DOI: 10.7554/eLife.04531.001

Introduction

Protein encoding regions of all organisms, eukaryotic, bacterial and viral, are subject to a number of functional constraints in addition to coding capacity, many of which contribute to regulation of translation. These include the widely reported biases in the relative frequencies of codons encoding the same amino acid (Bennetzen and Hall, 1982; Sharp et al., 2005; Wu et al., 2010) which in some organisms represents optimisation of the coding sequence for specific tRNAs, elongation rates and translation accuracy (reviewed in Gingold and Pilpel, 2011). There are, in addition, consistent underand over-representations of codon pairs (CPs) in all organisms (Yarus and Folley, 1985; Gutman and Hatfield, 1989; Boycheva et al., 2003; Moura et al., 2005; Tats et al., 2008) that have been proposed to influence gene expression through alterations in translation efficiency.

Because of its potential effect on gene expression, altering CP frequencies towards those that are disfavoured in their hosts has recently been advocated as a novel strategy to reduce RNA virus replication (Coleman et al., 2008; Wimmer et al., 2009; Mueller et al., 2010; Martrus et al., 2013; Yang et al., 2013; Le Nouen et al., 2014; Ni et al., 2014). This procedure potentially provides the means to produce a new generation of safer, non-reverting, live attenuated vaccines. Classically, virus genomes have been empirically attenuated by serial passage in tissue-culture leading to the accumulation of mutations. This lengthy, stochastic process produced attenuated virus vaccine strains which have produced major effects on human (eg., poliovirus—Sabin vaccines) and animal health (eg., eradication of Rinderpest using the Plowright vaccine). However, reversion to virulence by back-mutation of characteristically a small number of key, attenuating, mutations is a well-known problem. Novel strategies by which synonymous coding changes are introduced to modify codon usage (Mueller et al., 2010) (Martrus et al., 2013; Yang et al., 2013; Le Nouen et al., 2014; Ni et al., 2014) have the advantage...
that the resulting virus attenuation is dependent on a large number of mutations each of which only slightly reduce replicative fitness, but taken together produce significant attenuation with greatly enhanced genetic stability.

As one of the first examples, Coleman et al. (2008) generated synthetic poliovirus capsid gene sequences containing codon pairs that were specifically disfavoured in human coding sequences. These CP de-optimised sequences were inserted into an infectious cDNA clone of poliovirus. Virus generated from these mutants showed a remarkably attenuated replication phenotype attributed by the authors to impaired translation efficiency. Codon pair de-optimisation (CPD) has since been developed as a strategy for the production of a wide range of other live attenuated virus vaccines including influenza A virus (IAV), porcine reproductive and respiratory syndrome virus (PRRSV), human immunodeficiency virus type 1 (HIV-1) and respiratory syncytial virus (Mueller et al., 2010; Martrus et al., 2013; Yang et al., 2013; Le Nouen et al., 2014; Ni et al., 2014).

While altering translation efficiency through manipulation of codon or codon pair usage may attenuate virus replication, other virus compositional features may additionally contribute to replication phenotypes. One prominent compositional abnormality among RNA and small DNA viruses infecting mammals and plants is the marked suppression of the frequencies of CpG and UpA dinucleotides...
The functional basis for this suppression was recently demonstrated by the marked attenuating effect of artificially increasing the numbers of CpG and UpA dinucleotides in the genome of echovirus 7 (E7; (Atkinson et al., 2014)). We and others (Burns et al., 2009) have speculated that effects of CpG/UpA frequencies on virus replication may indeed account for, at least in part, the attenuating effect of selecting disfavoured codon pairs in CPD mutant of poliovirus and other candidate attenuated vaccines. Supporting this conjecture, regression analysis of the effects of numerous compositional variables in a range of codon- and codon pair deoptimised mutants on poliovirus replication demonstrated the primary effect of CpG and UpA frequencies on replication ability rather than alterations in codon or codon pair usage (Burns et al., 2009).

In the current study, we have used a variety of bioinformatic analyses to investigate the relationship between dinucleotide frequencies and codon pair usage. We have subsequently designed and assessed the replication phenotypes and fitness of mutants of E7 constructed in such a way that allows effects of codon pair and dinucleotide frequency alterations to be separately altered. The findings demonstrate the primary influence of CpG and UpA frequencies on virus replication that was independent of codon pair usage and translation efficiency.

### Results

**Virus attenuation, CP and dinucleotide frequencies**

Coding regions of poliovirus, IAV, PRRSV and HIV-1 have all been subjected to CP de-optimisation and effects on virus replication quantified (Coleman et al., 2008; Mueller et al., 2010; Martrus et al., 2013; Yang et al., 2013; Ni et al., 2014). Despite their diversity of replication and translation mechanisms, each showed a similar relationship between the extent of CPD and reduction in virus replication ability (Table 1). Typically, 10-fold or greater attenuation in cell culture required >12–15% replacement of WT genome with CPD sequences. It is notable, however, that for each virus, CPD

<table>
<thead>
<tr>
<th>Virus</th>
<th>Gene</th>
<th>Prop’n</th>
<th>CP bias</th>
<th>CpG</th>
<th>UpA</th>
<th>CP bias</th>
<th>CpG</th>
<th>UpA</th>
<th>Replication Reduction</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poliovirus</td>
<td>Capsid</td>
<td>14.8%</td>
<td>−0.03</td>
<td>0.52</td>
<td>0.75</td>
<td>−0.46</td>
<td>1.34</td>
<td>1.25</td>
<td>×25</td>
<td>Coleman et al., 2008</td>
</tr>
<tr>
<td></td>
<td>Capsid</td>
<td>25.9%</td>
<td>−0.03</td>
<td>0.54</td>
<td>0.75</td>
<td>−0.46</td>
<td>1.31</td>
<td>1.27</td>
<td>×400</td>
<td></td>
</tr>
<tr>
<td>Influenza A virus*</td>
<td>HAmin</td>
<td>Segs.4</td>
<td>11.4%</td>
<td>0.02</td>
<td>0.43</td>
<td>0.64</td>
<td>−0.42</td>
<td>1.65</td>
<td>1.11 ×3.5</td>
<td>Mueller et al., 2010</td>
</tr>
<tr>
<td></td>
<td>HA/NPmin</td>
<td>Segs.4,5</td>
<td>21.3%</td>
<td>0.02</td>
<td>0.44</td>
<td>0.55</td>
<td>−0.42</td>
<td>1.56</td>
<td>1.14 ×14</td>
<td></td>
</tr>
<tr>
<td>PR8*</td>
<td>Segs.1,4,5</td>
<td>29.1%</td>
<td>0.01</td>
<td>0.43</td>
<td>0.53</td>
<td>0.41</td>
<td>−0.41</td>
<td>1.55</td>
<td>1.07 ×35</td>
<td></td>
</tr>
<tr>
<td>HIV-1</td>
<td>A gag</td>
<td>4.6%</td>
<td>0.03</td>
<td>0.47</td>
<td>1.04</td>
<td>−0.43</td>
<td>1.43</td>
<td>1.25</td>
<td>×7</td>
<td>Martrus et al., 2013</td>
</tr>
<tr>
<td></td>
<td>B gag</td>
<td>4.7%</td>
<td>0.08</td>
<td>0</td>
<td>0.91</td>
<td>−0.37</td>
<td>1.22</td>
<td>1.15</td>
<td>×3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C gag</td>
<td>4.8%</td>
<td>0.03</td>
<td>0.31</td>
<td>1.00</td>
<td>−0.38</td>
<td>1.50</td>
<td>1.09</td>
<td>×8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D gag</td>
<td>2.1%</td>
<td>−0.02</td>
<td>0</td>
<td>0.49</td>
<td>−0.42</td>
<td>1.47</td>
<td>0.99</td>
<td>×1.5</td>
<td></td>
</tr>
<tr>
<td>PRRSV</td>
<td>SAVE5 gp5</td>
<td>2.6%†</td>
<td>−0.06</td>
<td>0.63</td>
<td>0.73</td>
<td>−0.38</td>
<td>1.37</td>
<td>1.14</td>
<td>×4‡</td>
<td>Ni et al., 2014</td>
</tr>
</tbody>
</table>

* Codon pair minimised sequences of IAV were not provided in (Coleman et al., 2008) and for the purposes of comparison these have been reconstructed in SSE. Note that the CP scores described in Table 1 of that paper (−0.386, −0.420 and −0.421 for PB1, HA and NP respectively) are not minimum scores; these are in fact −0.533, −0.585 and −0.602. Therefore, for the purposes of comparison, CP score minimisation in the current study was targeted to the former values. Although the sequences generated by SSE were not identical to those obtained previously, they would demonstrate a similar distortion of dinucleotide frequencies to those used in the previous study (Coleman et al., 2008).

† Mutated region only (positions 147–542 in gp5).

‡ Data from replication assay in PAM cells.
invariably increased frequencies of CpG and UpA dinucleotides (Table 1), typically from 0.4–0.6 to 1.4–1.6 (CpG) and from 0.5–0.8 to 1.1–1.4 (UpA) in the mutated regions.

This linkage can be accounted for at least in part by the association between CP choice and the identity of the dinucleotide between the third and first (3–1) codon positions. We analysed these parameters in coding regions of a curated dataset of over 35,170 human mRNA sequences. The representation of each of the 3904 codon pairs found in coding sequences (ie., 61 × 64) was calculated, taking into account both the nucleotide composition of the sequences and the amino acid usage as previously described (Gutman and Hatfield, 1989). Relative under- and over-representation of each was indicated in a heat map, with values ranging from −0.222 to +0.271 (mean 0.072). Values were plotted on x- and y-axes using values that reflected base identities at each of six positions in the codon pair (Figure 1A). Most of the 256 CPs with CpG at the 3–1 position (sixth main column) were markedly under-represented. There was further influence of codon pair position 6 (CP score was more suppressed if A or U) but with generally minimal and inconsistent influences of nucleotide identities at other codon positions (Bennetzen and Hall, 1982; Buchan et al., 2006; Atkinson et al., 2014). In the overall distribution of human codon pair representations, codon pairs containing CpG across the codon boundary distributed towards the negative tail of the distribution of CP score (Figure 1B) and accounted for almost all those with scores below −1.25.

CPs with UpA at the 3–1 position were also specifically under-represented in human mRNA sequences (Figure 1A,B), consistent with global under-representation of this dinucleotide in coding sequences from eukaryotes (Beutler et al., 1989; Duan and Antezana, 2003). The dataset additionally demonstrated over-representation of CpA and UpG dinucleotides at the 3–1 position; these are typically created by the (methylation-associated) C>T transition upstream of G (fifth and 14th main columns in Figure 1A) and of CpC and CpU (Simmen, 2008). However, with a few exceptions, such as the prominent over-representation of GCG|GCG and CCG|CCG, other codon pairs showed infrequent or minor differences in representation. The avoidance of CpG and UpA in human mRNA sequences at the 3–1 position was further manifested at other three codon position (Bennetzen and Hall, 1982; Atkinson et al., 2014); among the 61 degenerate codons, those containing CpG or UpA at these positions showed lower relative synonymous codon usage than

---

**Figure 1.** (A) CP usage in human coding sequences arranged in a 64 x 64 grid. CP frequencies relative to those expected from nucleotide and amino acid frequencies (CP bias) are colour coded in a heat map. The primary division on the x-axis is by identity of the 3–1 dinucleotide as annotated. Within these, further divisions within each of the 16 columns show the identity of the nucleotide at position 2 (A, C, G or U). The y-axis records nucleotides at positions 4 (4 main divisions on the y-axis), 1 (4 subdivisions of position 4) and 6 (4 subdivisions of position 1). Positions of unused codon pairs containing a 5′ stop codon (translated as *|x) are shaded in grey. CP usage heat maps for A. thaliana, C. elegans and E. coli coding sequences are shown in Figure 1—figure supplement 2A–2C. (B) Distribution of codon pair bias scores in human coding sequences; separate labelling of the 64 codon pairs with CpG (red) or UpA (blue) across the codon junction (3–1) demonstrates their consistent under-representation based on their component nucleotide and amino acid frequencies. The distribution of codon pair scores for A. thaliana, C. elegans and E. coli are shown in Figure 1. Continued on next page.
Figure 1. Continued

**Figure 1—figure supplement 3A–3C.** Correlations between codon pair scores between human coding sequences and those of A. thaliana, C. elegans and E. coli are shown in Figure 1—figure supplement 4. DOI: 10.7554/eLife.04531.004

The following figure supplements are available for figure 1:

**Figure supplement 1.** Distribution of relative synonymous codon usage values for degenerate codons in the human genome (stop codons were excluded).
DOI: 10.7554/eLife.04531.005

**Figure supplement 2.** CP scores of codon pairs of (A) A. thaliana, (B) C. elegans and (C) E. coli ORFeomes.
DOI: 10.7554/eLife.04531.006

**Figure supplement 3.** Distribution of codon pair scores for other organisms-(A) A. thaliana, (B) C. elegans and (C) E. coli, with separate representation of codon pairs with CpG and UpA across the codon junction.
DOI: 10.7554/eLife.04531.007

**Figure supplement 4.** Correlation between representations of human codon pairs (x-axis) with those of other organisms-(A) A. thaliana, (B) C. elegans and (C) E. coli (y-axis).
DOI: 10.7554/eLife.04531.008

Codon pairs was a simple consequence of avoiding disfavoured codon pairs. Conversely, it could be conceptualised that codon pair choice is driven in part through avoidance of specific dinucleotides. To resolve this functionally, we compared replication dynamics and relative fitness of native E7 with a series of novel mutants of E7 in which dinucleotide frequencies and CP usage were independently manipulated (Figure 2; Table 2). To achieve this, a mutational program was developed (Sequence Mutate in the SSE package (Simmonds, 2012) that allowed synonymous changes to be introduced into a coding sequence to achieve a pre-specified CP score target while under constraints such as retaining CpG and UpA dinucleotide frequencies and mononucleotide composition.

The mutant, Min-E was constructed from two genome regions, together comprising 31% of the E7 genome, in which the coding sequence possessed the minimum possible CP score (~0.111) while retaining identical CpG and UpA frequencies as WT virus (CP score: −0.014; CpG: 0.525; UpA: 0.718; Figure 2, Table 2). Inserts with the same CP frequencies as Min-E but without dinucleotide frequency constraints (Min-U; CpG: 0.82; UpA: 0.95) or where CpG and UpA frequencies were maximised (Min-H; CpG: 1.3; UpA: 0.98) were generated similarly. The three mutants provided the opportunity to investigate effects of dinucleotide frequency differences of viral fitness without the compounding effect of CP bias. It was similarly possible to compare fitness of the mutant, Max-U, with a maximised CP score (0.320) but with similar CpG and UpA frequencies to WT with the previously described mutant, cu|cu, with minimised CpG and UpA frequencies (0 and 0.22 respectively) but a CP score marginally greater that WT (0.11; Figure 2). P|P was the permuted mutant control with randomised codon order but identical coding and dinucleotide frequencies to WT sequence.

If CP usage solely determined virus replication ability, the seven mutants would be expected to display the following fitness ranking:

Max-U > cu|cu > (WT = P|P) > (Min-H = Min-U = Min-E).

Conversely, if virus fitness was determined by CpG and UpA dinucleotide frequencies, the following ranking would be expected:

cu|cu > Max-U > (WT = P|P = Min-E) > Min-U > Min-H.

These predictions were determined by generation and infectivity measurement of virus stocks corresponding to the seven mutants and comparing their relative fitness in competition and replication assays (Figures 3,4).
Replication phenotypes

Full length RNA transcribed from each E7 mutant cDNA constructs all generated infectious virus after transfection into RD cells. Stocks of virus were generated from WT and each mutant and infectivity quantified by quantal limiting dilution. To investigate replication kinetics, RD cells were infected with WT and each mutant at an MOI of 0.03 in triplicate and infectivity of supernatants measured at 8, 18 and 30 hr (Figure 3;). During the exponential period of replication (8 and 8 hr), Min-U and Min-H mutants showed 1 and >2 log reductions in virus replication respectively compared to WT E7. Contrastingly, the CpG/UpA-minimised mutant, cu|cu replicated to approximately 1 log higher levels that WT. Significantly for the analysis of effects of CP and dinucleotide frequencies on replication, virus titres obtained from mutant with identical (Min-E, CDLR) or similar (Max) CpG/UpA frequencies to WT were highly similar at both time points. At the last timepoint (30 hr), RD cells infected with WT, CDLR, Min-E, Max and cu|cu were entirely destroyed or almost entirely destroyed (Max-U) and showed similar residual infectivities, while those infected with Min-H showed an incomplete cytopathic effect.

Competition assays were used as a more stringent measure of fitness differences in mutants with different codon pair biases. Equal MOIs of WT and mutants were co-inoculated onto RD cells and serially passaged up to ten times. Population compositions were determined by amplification of sequences across modified regions and cleavage with restriction enzymes that differentiated WT mutant sequences from each other (Figure 4A,B; Table 3). In the examples of competition assays...
Max-U showed similar fitness to WT but a greater population representation at passage 10. cu|cu completely out-competed Max-U by passage 10 while in the final example, WT and Min-E showed equal fitness at passage 5 and at passage 10 (Figure 4B). A total of 12 pairwise comparisons were made and outcomes in terms of population representation recorded at passage 10 (Figure 4B; see Key). The results are internally consistent and with their replication kinetics (Figure 3) and indicate the following fitness ranking:

\[
\text{cu | cu} > \text{Max-U} > (\text{WT P | P = Min-E}) > \text{Min-U} > \text{Min-H}
\]

Using the Spearman rank correlation test, fitness ranking was significantly associated with CpG and UpA frequencies in the insert region (p < 0.001) but showed no association with CP frequencies and other measures of codon usage that potentially influence translation rates, codon adaptation index (CAI) and effective number of codons (ENc) (Table 4). Consistently, these results demonstrate that when altered independently from CP bias, only dinucleotide frequencies were associated with replication fitness.

### Comparison of translation efficiencies

The maxim that any effects of CP frequencies on replication are mediated through its influence on translation efficiency was investigated for the mutants constructed in the study. Translation assays were evaluated in vitro to avoid effects mediated through stress response-related RNA recognition mechanisms that restrict E7 translation and subsequent replication immediately after entry (Atkinson et al., 2014). Viral RNA transcripts from E7 WT and mutant cDNA clones were used to program rabbit reticulocyte lysates in the presence of \(^{35}S\)-methionine. Electrophoresis of reactions after 3 hr showed translation of several bands representing cleaved and partially cleaved E7 proteins (Figure 5; Figure 5—figure supplement 1). Translation efficiencies of each of the mutant E7 transcripts were comparable to WT RNA; what variability there was between mutants (Figure 5—figure supplement 1) did not correlate with replication fitness (R = −0.075; p > 0.5; Table 4). This indicates that, at least in a whole genome context, alteration of either CP or dinucleotide frequencies had no significant effect on viral polyprotein translation and therefore cannot be attributed to the marked differences in replication phenotypes observed.

---

**Table 2.** Composition and codon usage of E7 wt and mutant insert sequences

<table>
<thead>
<tr>
<th>Region</th>
<th>Sequence (Symbol)</th>
<th>G+C content</th>
<th>CpG Total*</th>
<th>O/E ratio†,‡</th>
<th>UpA Total*</th>
<th>O/E ratio†,‡</th>
<th>Codon Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CAI¶</td>
</tr>
<tr>
<td>1</td>
<td>Native (WT)</td>
<td>47.6%</td>
<td>51 (−−)</td>
<td>0.730</td>
<td>62 (−−)</td>
<td>0.742</td>
<td>0.043</td>
</tr>
<tr>
<td></td>
<td>Permuted (P)</td>
<td>47.6%</td>
<td>51 (0)</td>
<td>0.730</td>
<td>2 (0)</td>
<td>0.742</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>CpG/UpAL (cu)</td>
<td>47.5%</td>
<td>0 (−51)</td>
<td>0</td>
<td>19 (−43)</td>
<td>0.227</td>
<td>0.087</td>
</tr>
<tr>
<td></td>
<td>Max-U</td>
<td>50.1%</td>
<td>47 (−4)</td>
<td>0.610</td>
<td>43 (−19)</td>
<td>0.573</td>
<td>0.328</td>
</tr>
<tr>
<td></td>
<td>Min_E</td>
<td>47.5%</td>
<td>51 (0)</td>
<td>0.736</td>
<td>62 (0)</td>
<td>0.735</td>
<td>0.131</td>
</tr>
<tr>
<td></td>
<td>Min_U</td>
<td>47.5%</td>
<td>69 (+18)</td>
<td>0.992</td>
<td>76 (+14)</td>
<td>0.939</td>
<td>0.134</td>
</tr>
<tr>
<td></td>
<td>Min_H</td>
<td>49.8%</td>
<td>106 (+55)</td>
<td>1.400</td>
<td>79 (+17)</td>
<td>0.981</td>
<td>0.130</td>
</tr>
<tr>
<td>2</td>
<td>Native (WT)</td>
<td>47.1%</td>
<td>18 (−−)</td>
<td>0.320</td>
<td>48 (−−)</td>
<td>0.695</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>Permuted (P)</td>
<td>47.6%</td>
<td>18 (0)</td>
<td>0.320</td>
<td>48 (0)</td>
<td>0.695</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>CpG/UpAL (cu)</td>
<td>48.5%</td>
<td>0 (−18)</td>
<td>0</td>
<td>48 (0)</td>
<td>0.214</td>
<td>0.118</td>
</tr>
<tr>
<td></td>
<td>Max-U</td>
<td>46.3%</td>
<td>24 (+6)</td>
<td>0.440</td>
<td>43 (−3)</td>
<td>0.601</td>
<td>0.311</td>
</tr>
<tr>
<td></td>
<td>Min_E</td>
<td>45.7%</td>
<td>18 (0)</td>
<td>0.343</td>
<td>48 (0)</td>
<td>0.657</td>
<td>0.091</td>
</tr>
<tr>
<td></td>
<td>Min_U</td>
<td>47.4%</td>
<td>37 (+19)</td>
<td>0.649</td>
<td>50 (+2)</td>
<td>0.738</td>
<td>0.083</td>
</tr>
<tr>
<td></td>
<td>Min-H</td>
<td>47.8%</td>
<td>68 (+50)</td>
<td>1.172</td>
<td>65 (+15)</td>
<td>0.970</td>
<td>0.085</td>
</tr>
</tbody>
</table>

*Total number of CpG and UpA dinucleotides in sequence. Changes in numbers between mutated and original WT sequence are indicated in parentheses.
†Ratio of observed dinucleotide frequency (O) to that expected based on mononucleotide composition (E) that is, f(CpG)/f(C) × f(G).
‡Values deliberately changed are shown in red (maximised) and blue (minimised).
¶Calculated from http://genomes.urv.es/CAIcal/ (Puigbo et al., 2008).
This study sought to disentangle effects of codon pair usage and nucleotide frequencies in a re-examination of their effects on the replication of an RNA virus, E7. In the literature, studies have documented effects of CP de-optimisation on virus attenuation without reference to effects of this procedure on dinucleotide frequencies (Coleman et al., 2008; Mueller et al., 2010; Martrus et al., 2013; Yang et al., 2013; Le Nouen et al., 2014; Ni et al., 2014). While frequencies of both of these dinucleotides are suppressed in most classes of mammalian RNA viruses (Rima and McFerran, 1997), all sequences modified to select disfavoured CPs (Coleman et al., 2008; Mueller et al., 2010; Martrus et al., 2013; Yang et al., 2013; Ni et al., 2014) consistently elevated frequencies of CpG and UpA dinucleotides to levels to 2.5–threefold higher levels than the original native sequences (Table 1). As documented in other studies (Burns et al., 2009; Atkinson et al., 2014), these dinucleotide frequencies may contribute additionally to the observed attenuation of virus replication.

Through construction of mutants of E7 in which CP frequencies was altered while keeping dinucleotide frequencies constant (WT/Min-E) and conversely, generating viruses with the same or similar CP biases but different dinucleotide frequencies (eg., Min-E/Min-U/Min-U and cu|cu/WT/Max-U), we were able to separate potential influences of these compositional variables on replication phenotype and fitness. The fitness ranking derived from competition assays (cu|cu > Max > (WT = P|P = Min-E) > Min-U > Min-H) demonstrated that it was dinucleotide frequencies that significantly influenced fitness while differences in CP usage showed no detectable phenotypic effect. Moreover, if CPD were to influence virus replication then its effect would be manifested through changes in translation rate; however, no measurable differences in translation efficiency were detected between WT and CP-optimised (Max) or de-optimised (Min) template RNA sequences. These findings are broadly consistent with results

**Figure 3.** Replication of WT and mutants of E7 with altered CP and dinucleotide frequencies. Bars are shaded diagrammatically based on their relative CpG/UpA composition. RD cells were infected with E7 WT, at an MOI of 0.03 and infectious titres quantified at 8, 18 and 30 hr time points post inoculation (p.i.) by TCID$_{50}$ determination. Results are the mean of three biological replicates; error bars show standard errors of the mean.

DOI: 10.7554/eLife.04531.011

**Discussion**

This study sought to disentangle effects of codon pair usage and nucleotide frequencies in a re-examination of their effects on the replication of an RNA virus, E7. In the literature, studies have documented effects of CP de-optimisation on virus attenuation without reference to effects of this procedure on dinucleotide frequencies (Coleman et al., 2008; Mueller et al., 2010; Martrus et al., 2013; Yang et al., 2013; Le Nouen et al., 2014; Ni et al., 2014). While frequencies of both of these dinucleotides are suppressed in most classes of mammalian RNA viruses (Rima and McFerran, 1997), all sequences modified to select disfavoured CPs (Coleman et al., 2008; Mueller et al., 2010; Martrus et al., 2013; Yang et al., 2013; Ni et al., 2014) consistently elevated frequencies of CpG and UpA dinucleotides to levels to 2.5–threefold higher levels than the original native sequences (Table 1). As documented in other studies (Burns et al., 2009; Atkinson et al., 2014), these dinucleotide frequencies may contribute additionally to the observed attenuation of virus replication.

Through construction of mutants of E7 in which CP frequencies was altered while keeping dinucleotide frequencies constant (WT/Min-E) and conversely, generating viruses with the same or similar CP biases but different dinucleotide frequencies (eg., Min-E/Min-U/Min-U and cu|cu/WT/Max-U), we were able to separate potential influences of these compositional variables on replication phenotype and fitness. The fitness ranking derived from competition assays (cu|cu > Max > (WT = P|P = Min-E) > Min-U > Min-H) demonstrated that it was dinucleotide frequencies that significantly influenced fitness while differences in CP usage showed no detectable phenotypic effect. Moreover, if CPD were to influence virus replication then its effect would be manifested through changes in translation rate; however, no measurable differences in translation efficiency were detected between WT and CP-optimised (Max) or de-optimised (Min) template RNA sequences. These findings are broadly consistent with results
of previous translation assays of poliovirus mutants differing in CP choice in which relatively small differences in translation of PV-Min, WT and PV-Max mutants were clearly incompatible with the marked differences in their replicative ability (Coleman et al., 2008).

In a broader context, the finding that alterations in codon pair frequencies has no independent effect on virus replication are consistent with current understanding of the nature and driving forces behind codon pair usage in other organisms. Most importantly, there is no evidence that disfavoured CPs in eukaryotes, archaea or prokaryotes are those that specifically retard translation rates. Indeed, where specifically investigated, the opposite was observed. mRNA templates containing disfavoured codon pairs in *E. coli* were translated faster than those containing over-represented CPs (Irwin et al., 1995). The 16 codon pairs identified as most retarding translation of the *E. coli* his operon leader peptide gene has codon pair scores ranging from −0.94 to +0.54 and distributed around the centre of the distribution of codon pair scores (Figure 1—Figure Supplement 3) (Chevance et al., 2014).

The current consensus view is that CP usage in prokaryotes is governed functionally as a means to regulate gene expression rather than to maximise translation (Folley and Yarus, 1989; Irwin et al., 1995; Boycheva et al., 2003; Buchan et al., 2006).
In eukaryotic genomes, other factors underlie codon pair representation since coding regions in mRNA sequences and in non-transcribed genomic DNA showed similar biases in codon pair frequencies (Moura et al., 2007). CP frequencies must therefore be substantially determined by mutational events operating on DNA such as methylation and specific context-dependent errors during genome replication instead of any kind of optimisation or regulation of translation. The consistent under-representation of codon pairs with CpG in the 3–1 position (Figure 1; Buchan et al., 2006) in mammalian genomes indeed likely originates from DNA methylation-induced mutations in the nucleus. Our data showing similar rates of translation of Min-H and Max-U that show major differences in frequencies of CpG-containing CPs (Figure 5) are consistent with this interpretation.

Finally, there is no theoretical basis for the assumption that CPs are disfavoured because of their negative effects on translation efficiency and this concept runs counter to our growing understanding of the intricate mechanisms that govern gene expression. In all organisms, coding sequences differ in codon usage, match to tRNA abundances, mRNA stability and initiation sites to regulate rates and fidelity of protein expression (reviewed in Gingold and Pilpel, 2011). Some of the variability in CP usage observed in the three domains of life (Moura et al., 2005; Sharp et al., 2005; Tats et al., 2008; Wang and Li, 2009) likely represents aspects of that control, rather than as a means to simply maximise translation.

**Mechanism of attenuation**

Understanding what limits the replication of viruses with altered CP and dinucleotide frequencies is critical in the evaluation of their broader safety as attenuated virus vaccines. The proposed mechanism in which alterations in CP bias alter translation efficiency and it is this that inhibits virus replication introduces a conceptual model in which it is the virus that is intrinsically defective. With the large number of mutations required for reversion, such viruses should be stably attenuated in whatever context they are used. However, as we have now shown, the replication defect of CPD viruses is actually mediated through alterations in dinucleotide frequencies in the genome that influence their recognition by the cell. In this alternative paradigm, viruses with elevated frequencies of CpG and UpA are not intrinsically defective but they are more readily recognised by the cell and prevented from initiating replication. Their attenuation is therefore dependent on the efficacy of the host innate immune response.

**Table 3. Enzymes used in selective digests for competition ASSAYS**

<table>
<thead>
<tr>
<th>Virus 1</th>
<th>Virus 2</th>
<th>Region</th>
<th>Enzyme</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>W</td>
<td>Min-H</td>
<td>1</td>
<td>Spe</td>
</tr>
<tr>
<td>W</td>
<td>W</td>
<td>Min-H</td>
<td>1</td>
<td>Sacl</td>
</tr>
<tr>
<td>W</td>
<td>W</td>
<td>Min-E</td>
<td>1</td>
<td>Ncol</td>
</tr>
<tr>
<td>W</td>
<td>W</td>
<td>Min-E</td>
<td>1</td>
<td>Ncol</td>
</tr>
<tr>
<td>W</td>
<td>W</td>
<td>cu</td>
<td>cu</td>
<td>1</td>
</tr>
<tr>
<td>P</td>
<td>P</td>
<td>cu</td>
<td>cu</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table 4. Correlation between fitness ranking and sequence composition**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Spearman R</th>
<th>p value†</th>
</tr>
</thead>
<tbody>
<tr>
<td>CpG/UpA*</td>
<td>1.0</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>CP bias</td>
<td>−0.70</td>
<td>0.1 (n.s.)</td>
</tr>
<tr>
<td>CAI</td>
<td>−0.334</td>
<td>&gt;0.5 (n.s.)</td>
</tr>
<tr>
<td>ENc</td>
<td>0.593</td>
<td>&gt;0.5 (n.s.)</td>
</tr>
<tr>
<td>G + C content</td>
<td>0.075</td>
<td>&gt;0.5 (n.s.)</td>
</tr>
<tr>
<td>Translation efficiency</td>
<td>−0.074</td>
<td>&gt;0.5 (n.s.)</td>
</tr>
</tbody>
</table>

*Number of CpG and UpA dinucleotides in insert region. †From values tabulated in (Ramsey, 1989). n.s.: not significant. DOI: 10.7554/eLife.04531.014

DOI: 10.7554/eLife.04531.013

DOI: 10.7554/eLife.04531.015
The cellular mechanisms responsible for differential recognition and response to RNA sequences with different dinucleotide composition are currently unknown. In our previous study, we obtained evidence that replication inhibition of high CpG/UpA mutants of E7 occurred shortly after cell entry and was not mediated through conventional pattern recognition receptors (Atkinson et al., 2014). In that study, we additionally demonstrated that it was additionally not the result of differences between high and low CpG/UpA viruses in their sensitivity to the cellular interferon response. We did observe, however, that the attenuated phenotype of high mutants could be entirely reversed by the kinase inhibitor, C16, a finding that suggests that recognition may occur through an as yet uncharacterised PKR-related component of the stress response pathway in the cell.

Both the adaptive and innate arms of the human immune system are highly polymorphic with remarkable variability in function and expression of many key components of recognition or effector proteins mediating antiviral responses (Thomas et al., 2009; Everitt et al., 2012; Hambleton et al., 2013; Pothlichet and Quintana-Murci, 2013). Although uncharacterised mechanistically, there is clearly a potential danger that pathways that restrict the replication of high CpG/UpA RNA viruses may be similarly variable in the efficacy in humans and in veterinary species with different genetic backgrounds. The attenuation of live vaccines and safety margins established for their large scale use may be similarly variable; investigation of population differences in innate cellular responses to viruses of different dinucleotide compositions is essential in the evaluation of the safety of this new generation of high CpG/UpA live attenuated vaccines.

Materials and methods

Cell culture and cell lines

RNA transcripts of the pT7:E7 infectious cDNA clone of the isolate Wallace (accession number AF465516) were used to generate E7 viral stocks. E7 was propagated in rhabdomyosarcoma (RD) cells using Dulbecco modified Eagle medium (DMEM) with 10% foetal calf serum (FCS), penicillin (100 U/ml) and streptomycin (100 µg/ml). All cells were maintained at 37°C with 5% CO₂.

In silico design of CpG and UpA modified viruses

The two regions in the E7 genome used previously to investigate effects of dinucleotide frequencies on virus replication (Atkinson et al., 2014) were used in the current study (Region 1: 1878–3119 and 5403–6462). Previously characterised mutants comprised the CpG/UpA-low mutant cu|cu with all CpG dinucleotides and as many UpA dinucleotides possible eliminated and the permuted mutant P|P in which codon order was permuted while retaining protein coding and native dinucleotide frequencies. Further mutants (Max-U, Min-E, Min-U, Min-H) are described in the main text; sequences listed in Supplementary file 1.

Bioinformatics analysis

Manipulation of dinucleotide frequencies and codon pair scores in coding sequences was performed using the program Sequence Mutate in version 1.2 of the SSE package (Simmonds, 2012). Reference
datasets of human, A. thaliana, C. elegans and E. coli messenger RNA sequences were obtained from the Refseq database (http://www.ncbi.nlm.nih.gov/nuccore). Codon pair usage tables were generated from coding regions of each mRNA sequence datasets from the four organisms using the program Composition Scan in the SSE package (Simmonds, 2012). Codon pair tables generated by SSE were used to calculate CP frequencies for all mRNA sequences with coding regions >200 bases in length from each organism. These comprised 35,770 human mRNA sequences, 32,768 from A. thaliana, 24,093 from C. elegans and 4316 from E. coli.

The codon pair table generated by SSE from our dataset of human mRNA sequences was used in preference to that previously described (Table S2 in reference Coleman et al., 2008) because of the larger number of human mRNA sequences now available. The previously published dataset as presented additionally unaccountably omitted a large number of codon pairs (3586 were listed instead of the 3904 expected—61 x 64). There was however a good correlation between CP frequencies between the two datasets (Figure 6).

The codon adaptation index for human codon usage was calculated through the website http://genomes.urv.es/CAIcal/ (Puigbo et al., 2008). The effective number of codons (Enc; Wright, 1990) and CP usage (Buchan et al., 2006; Coleman et al., 2008) were calculated using the program Composition Scan in SSE.

**Clone construction and recovery of mutant viruses**

Mutant E7 constructs with altered CP frequencies were generated from custom synthesised DNA sequences (Eurofins Genomics, Ebersberg, Germany). Mutant clones were constructed as previously described (Atkinson et al., 2014). All clones were sequenced over the insert regions prior to further applications. Infectious virus from each cDNA clone was recovered by transfection of RNA transcripts produced from plasmids linearised using NotI using a Riboprobe System-T7 in vitro transcription kit.
(Promega Ltd. Southampton, UK). 100 ng of RNA was transfected into RD cells using Lipofectamine 2000 (Invitrogen, Life Technologies Ltd., Paisley, UK) according to the manufacturer’s instructions. The resulting cell lysates were used to generate passage 1 stocks by re-infecting RD cells. Viral titres were determined by TCID₅₀ titration in RD cells.

**Replication phenotype**

Multi-step growth curves for each virus were generated by infecting RD cells in triplicate in 24-well plates at an MOI of 0.03 as previously described (Atkinson et al., 2014). Supernatant collected at time points (8, 18 and 30 hr post-infection) were assayed for infectivity by quantal dilution. Competition assays were performed as previously described. Briefly, equal titres of virus pairs (combined MOI = 0.01) were applied simultaneously to RD cells in 25 cm² bottles. Following the development of CPE, supernatant was collected and 300 µl applied to fresh RD cells. This was continued for up to 10 passages. The results of the competition assays were determined by restriction enzyme digestion of the amplicon amplified from Region 1 by combined reverse transcription—PCR (Atkinson et al., 2014). Restriction enzymes used to differentiate each mutant pair are listed in Table 3.

**In vitro transcription and translation**

RNAs were produced by in vitro T7 transcription (Riboprobe System T7, Promega) of the various cDNA plasmids, each linearised with NotI (Promega). Transcript RNAs were used to program nuclease-treated rabbit reticulocyte lysates (Promega) supplemented with HeLa cell S10 cytoplasmic extracts (Dundee Cell Products, Dundee, UK). Reactions were set-up as follows; 7 µl rabbit reticulocyte lysate, transcript RNA (0.25–2 µg), 0.5 µl 1 mM amino acid mix (minus methionine), 0.5 µl [³⁵S]-methionine (1200 Ci/mmol), 10 U RNasin Ribonuclease Inhibitor and 2.25 µl HeLa cell extract in a total volume of 12.5 µl. Reactions were incubated at 30°C for 3 hr and analysed by SDS-PAGE (4–20% Tris-Glycine, Expeideon Ltd. Cambridge, UK). Gels were exposed to film (Thermo Scientific, Basingstoke, UK) for 1–4 days at −70°C. To determine the relative density of the protein bands, densitometry was carried out on the scanned gel image using ImageJ 1.48 software (http://imagej.nih.gov/ij).

**Acknowledgements**

The authors would like to thank John Nicholson and Ashley Pearson for technical assistance with the cloning and recovery of mutants of E7. The study was funded by a project grant from the Wellcome Trust (WT087628MA) and the BBSRC (BB/K003801/1 and BB/L004526/1).

**Additional information**

**Funding**

<table>
<thead>
<tr>
<th>Funder</th>
<th>Grant reference number</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wellcome Trust</td>
<td>WT087628MA</td>
<td>David J Evans, Peter Simmonds</td>
</tr>
<tr>
<td>Biotechnology and Biological Sciences Research Council</td>
<td>BB/K003801/1</td>
<td>Martin D Ryan</td>
</tr>
<tr>
<td>Biotechnology and Biological Sciences Research Council</td>
<td>BB/L004526/1</td>
<td>Martin D Ryan</td>
</tr>
</tbody>
</table>

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

**Author contributions**

FT, Acquisition of data, Analysis and interpretation of data; NJA, Conception and design, Acquisition of data, Analysis and interpretation of data; DJE, MDR, PS, Conception and design, Analysis and interpretation of data, Drafting and revising the article

**Author ORCIDs**

Fiona Tulloch, [http://orcid.org/0000-0003-0859-2867](http://orcid.org/0000-0003-0859-2867)
Additional files

Supplementary file
• Supplementary file 1. Nucleotide sequences of mutants used in the study. Supplementary file 1.docx http://dx.doi.org/10.7488/ds/188.

DOI: 10.7554/eLife.04531.019

Major datasets

The following datasets were generated:

<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Year</th>
<th>Dataset title</th>
<th>Dataset ID and/or URL</th>
<th>Database, license, and accessibility information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simmonds P</td>
<td>2014</td>
<td>Composition and codon usage metrics for human mRNA</td>
<td><a href="http://dx.doi.org/10.7488/ds/188">http://dx.doi.org/10.7488/ds/188</a></td>
<td>All data is licensed under a Creative Commons Attribution 4.0 International License.</td>
</tr>
<tr>
<td>Simmonds P</td>
<td>2014</td>
<td>Composition and codon usage metrics for A. thaliana mRNA</td>
<td><a href="http://dx.doi.org/10.7488/ds/188">http://dx.doi.org/10.7488/ds/188</a></td>
<td>All data is licensed under a Creative Commons Attribution 4.0 International License.</td>
</tr>
<tr>
<td>Simmonds P</td>
<td>2014</td>
<td>Composition and codon usage metrics for C. elegans mRNA</td>
<td><a href="http://dx.doi.org/10.7488/ds/188">http://dx.doi.org/10.7488/ds/188</a></td>
<td>All data is licensed under a Creative Commons Attribution 4.0 International License.</td>
</tr>
<tr>
<td>Simmonds P</td>
<td>2014</td>
<td>Composition and codon usage metrics for E. coli mRNA</td>
<td><a href="http://dx.doi.org/10.7488/ds/188">http://dx.doi.org/10.7488/ds/188</a></td>
<td>All data is licensed under a Creative Commons Attribution 4.0 International License.</td>
</tr>
<tr>
<td>Simmonds P</td>
<td>2014</td>
<td>Codon pair usage tables for human, A. thaliana, C. elegans and E. coli</td>
<td><a href="http://dx.doi.org/10.7488/ds/188">http://dx.doi.org/10.7488/ds/188</a></td>
<td>All data is licensed under a Creative Commons Attribution 4.0 International License.</td>
</tr>
</tbody>
</table>

References


