Positive selection analysis of overlapping reading frames is invalid

Christopher Monit[1], Richard A. Goldstein[1], Greg Towers[1] and Stéphane Hué[2]

1. Division of Infection and Immunity, University College London, London, UK
2. Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK

c.monit.12@ucl.ac.uk; r.goldstein@ucl.ac.uk; g.towers@ucl.ac.uk; stephane.hue@lshtm.ac.uk

Dear Sir,

In a recent *ARHR* publication by Roy *et al.* (*Intersubtype Genetic Variation of HIV-1 Tat Exon 1*, available ahead of print), the authors report the identification of 30 codon sites under positive selection in the first exon of the HIV-1 *tat* gene. Unfortunately, the authors have not considered the presence of overlapping coding sequences in HIV-1, invalidating most of their positive selection analysis.

Conventional phylogenetic selection analyses compare the rate of synonymous codon substitution (*dS*) with the rate of non-synonymous codon substitution (*dN*). A ratio of *dN*/*dS* exceeding 1 indicates positive selection, as the rate of amino acid change is greater than the rate of (ostensibly) neutral evolution. Selection acting on overlapping coding sequences can invalidate such analyses, as non-synonymous sites in one overlapping reading frame will correspond to a synonymous site in another[1].

In the HIV-1 genome, the first seven codons at the 5' end of *tat* exon 1 overlap with the 3' end of the *vpr* gene, while the last 26 codons overlap with *rev* exon 1 (HXB2 reference sequence; http://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html). Thus, the first nucleotide position in *vpr*/*rev* codons corresponds to the third nucleotide position in *tat* codons, as shown in Figure 1. Substitutions in the first nucleotide positions in the *vpr*/*rev* codons are likely to be non-synonymous, meaning purifying selection acting on *vpr* and *rev* would reduce the substitution rate at these positions. This corresponds to a decreased substitution rate at the third nucleotide position in the *tat* codons, which are likely to be synonymous, therefore inflating the *dN/dS* ratio for *tat* and resulting in the mistaken impression of positive selection. For this reason, the analysis described by Roy

et al. is only suitable for the region of *tat* between codons 8 and 46 that are not overlapping with other genes. Restricting the analysis to these residues eliminates the majority (19 of 30) of the sites identified by Roy *et al*. as under positive selection.

A number of approaches to studying selection specifically in overlapping coding sequences have been developed[1–3], though none are practical for routine analyses. Investigators undertaking phylogenetic selection analyses with conventional codon models must be aware of the bias introduced by ignoring overlapping coding sequences in virus or bacterial genomes. The scope of such work should be limited to regions of genes where there is no overlap.

Yours faithfully,

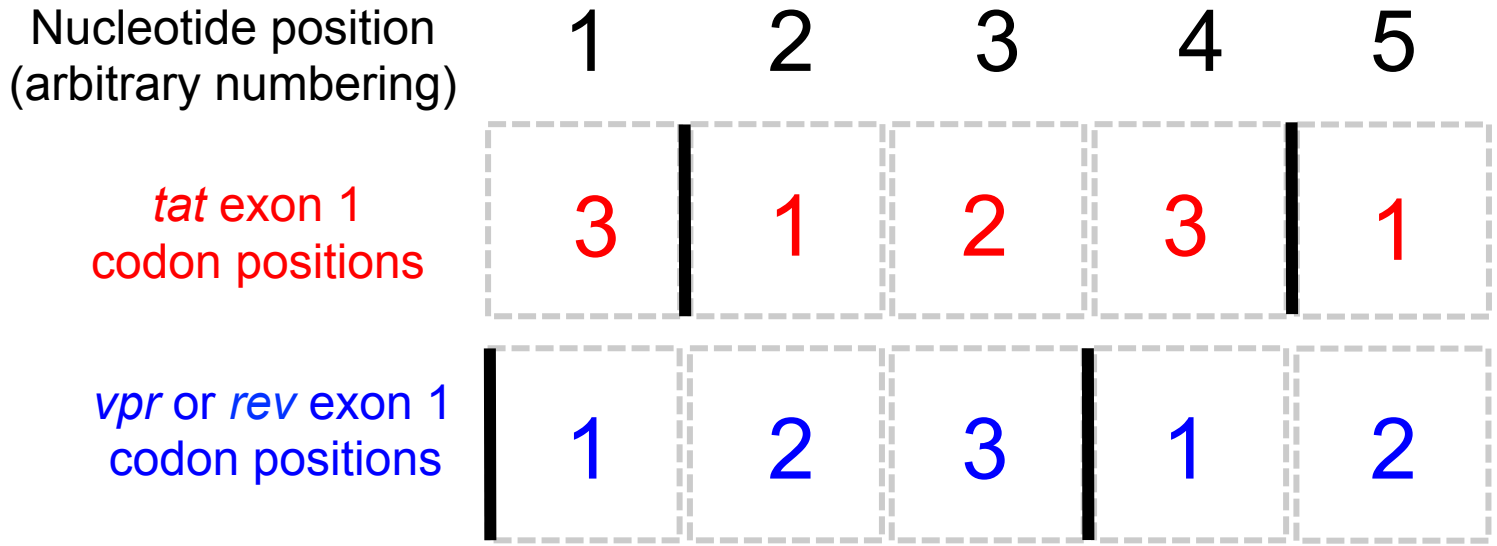
Christopher Monit, Richard A. Goldstein, Greg Towers and Stéphane Hué

References

1. Sabath, N., Landan, G. & Graur, D. A method for the simultaneous estimation of selection intensities in overlapping genes. *PLoS One* **3,** e3996 (2008).

2. Hein, J. & Stovlbaek, J. A maximum-likelihood approach to analyzing nonoverlapping and overlapping reading frames. *J. Mol. Evol.* **40,** 181–189 (1995).

3. Pedersen, A. M. & Jensen, J. L. A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Mol. Biol. Evol.* **18,** 763–76 (2001).


Nucleotide position
(arbitrary numbering)
1
2
3
4
5
tat exon 1
codon positions
3
1
2
3
1
vpr or rev exon 1
codon positions
1
2
3
1
2