Detecting overlapping coding sequences in virus genomes

Some additional information for Firth AE, Brown CM (2006) Detecting overlapping coding sequences in virus genomes. *BMC Bioinformatics* 7:75.

Caveats

For each component of the probability sum, only a single nucleotide and the single codon in each of the null and alternative model CDSs (if relevant) containing that nucleotide are considered; longer-range interactions are ignored. Furthermore, as far as amino acid and codon weightings are concerned, only the start and end points of the unknown mutation pathway connecting an aligned codon pair in S_1 and S_2 are considered. These simplifications have been discussed and justified in our previous paper [1].

Although the log(LR) statistic provides a useful means to discriminate between the null and alternative models, it can not be converted into a probability ratio, via $\frac{P(\text{ alternative model })}{P(\text{ null model })} \sim \exp(\log(\text{LR}))$. This is because (a) the null and alternative models can not be expected to have equal Bayesian prior probabilities; (b) the individual likelihoods (i.e. for a given nucleotide position in a given aligned sequence pair) are not all independent: adjacent nucleotide positions can be linked by being in the same codon, distant nucleotide positions can be linked by protein or RNA secondary structure interactions, and nucleotides at the same position in different sequence pairs can be linked by similar selective pressure on an encoded protein amino acid position; and (c) the mutation models are not entirely accurate.

Note that MLOGD can only detect functionally constrained regions of CDSs. Sometimes only part of a CDS is (strongly) functionally constrained. This can be particularly the case within CDS overlaps. Some CDS overlaps – especially the very short overlaps common in prokaryote genomes – may be purely regulatory in function, with the actual sequence in the overlap region being largely irrelevant. Some short ORFs in viruses may also be translated as part of a regulatory process, with the actual translation products being functionless.

Tests on known CDSs and non-coding ORFs in virus genomes

The software has been previously tested on overlapping CDSs in the Hepatitis B Virus and *Escherichia coli* genomes [1]. A database of results for the following virus groups is available on our website: Hepatitis B Virus (HBV), Avihepadnavirus (AHBV), Luteovirus, Polerovirus, Human Immunodeficiency Virus 1 (HIV-1), Hepatitis C Virus (HCV), Hepatitis G Virus (HGV), Enterovirus, Parechovirus, Dengue Virus, Japanese Encephalitis Virus (JEV), Tick-borne Encephalitis Virus (TBEV), Yellow Fever Virus (YFV) and Pestivirus (database now extended to 630 virus species).

For each of these 14 virus groups, a well-annotated sequence was chosen as a 'reference sequence'. All 37 known CDSs in these reference sequences were detected (i.e. positive $\sum_{\text{tree}} \log(\text{LR})$ statistics; Figure 1). Included within these are five examples of overlapping CDSs that are completely contained within other CDSs (one in HBV, one in AHBV, one in Luteovirus, one in Polerovirus and one in HIV-1), all of which have strongly positive $\sum_{\text{tree}} \log(\text{LR})$ statistics. The 20 CDSs that partially overlap other CDSs generally received a positive $\sum_{\text{tree}} \log(\text{LR})$ statistic within, as well as outside, the overlap region(s). Sometimes – especially when the overlap region is a relatively small fraction of the CDS – the $\sum_{\text{tree}} \log(\text{LR})$ statistic is negative within the overlap, indicating that the overlap region is not, on average, strongly functionally constrained in these CDSs. Here, the overlap may be more involved in regulating gene expression rather than in minimizing genome size.

Also included in the database, are the $\sum_{\text{tree}} \log(\text{LR})$ statistics for all start-stop ORFs of at least 40 codons in the reference sequences which were not already annotated as coding. Out of 404 ORFs, 252 had negative $\sum_{\text{tree}} \log(\text{LR})$ statistics (i.e. probably non-coding), while 152 had positive $\sum_{\text{tree}} \log(\text{LR})$ statistics (Figure 1). Of these, 139 were overlapping a known CDS in the -2 frame and had relatively low

per nt $\sum_{\text{tree}} \log(\text{LR})$ statistics. The -2 frame is particularly susceptible to false positives (see [1] or website for details), so in general such false positives should be discarded (the MLOGD software returns a warning message whenever there is CDS overlap in the -2 frame). In single-stranded RNA (ssRNA) positive-strand viruses, at least, all reverse read-frame ORFs may automatically be discarded. The remaining 13 positives all have very low per nt $\sum_{\text{tree}} \log(\text{LR})$ statistics relative to the per nt $\sum_{\text{tree}} \lambda$ – outside the range observed for the known CDSs (Figure 1). In addition 7 of the 13 have several stops introduced into the non-reference sequences. It is likely, therefore, that these ORFs merely represent the upper tail of random scatter of scores for non-coding ORFs.

Sensitivity as a function of query ORF length and input alignment divergence

The query ORF length and input alignment divergence may be combined into the single parameter $\sum_{\text{tree}} \Lambda$, i.e. the total number of mutations across the input alignment within the query region. The false positive and false negative rates, as a function of $\sum_{\text{tree}} \Lambda$, may be estimated by taking known non-coding, single-coding and double-coding regions of varying lengths, and seeing what fraction are correctly identified by the MLOGD statistic (see also [1] for more extensive sensitivity tests using simulated sequence data, and also tests of the effect of sequencing errors on the results).

Known single-coding regions were extracted from HCV, HGV, Enterovirus, Parechovirus, Dengue Virus, JEV, TBEV, YFV and Pestivirus alignments. Known double-coding regions were extracted from HBV, AHBV, Luteovirus, Polerovirus and HIV-1 alignments. Non-coding regions were crudely simulated by randomizing the columns in the HCV, HGV, Enterovirus, Parechovirus, Dengue Virus, JEV, TBEV, YFV and Pestivirus alignments. Many short sub-regions (10–320 codons in length) were extracted from these alignments. The single-coding regions were tested for coding potential in each of the six possible read-frames (null model = non-coding; alternative model = single-coding; the true read-frame is $+0$). The same regions were also tested for double-coding potential in the $+1$, $+2$, -0, -1 and -2 frames (null model = coding in the $+0$ frame; alternative model = also coding in another frame; see [1] or website for explanation of frame notation). The double-coding regions were tested for double-coding potential (null model = single-coding in one of the true read-frames; alternative model = double-coding). The non-coding regions were tested for coding potential in any of the six possible read-frames.

Figure 2 shows the raw $\sum_{\text{tree}} \log(\text{LR})$ (y-axis) versus $\sum_{\text{tree}} \Lambda$ (x-axis) statistics for a small random sample of these sub-alignments, for each null model – alternative model combination. In most cases, the line $y = 0$ is a close-to-optimal separator between the null and alternative models. For example, given a region that is actually single-coding in the $+0$ frame, if the null model is that the region is non-coding, then alternative models for coding in the $+1$, $+2$, -0 and -1 frames generally give negative $\sum_{\text{tree}} \log(\text{LR})$ statistics, while the alternative model for coding in the $+0$ frame generally gives positive $\sum_{\text{tree}} \log(\text{LR})$ statistics. Similarly, a region that is actually double-coding, generally receives a positive $\sum_{\text{tree}} \log(\text{LR})$ statistic if the alternative model is for double-coding in the correct frames, while a region that is actually single-coding generally receives a negative $\sum_{\text{tree}} \log(\text{LR})$ statistic if the alternative model is for double-coding in the $+1$, $+2$, -0 or -1 frame.

The -2 frame consistently produces false positives (for alternative model = double-coding or single-coding in the -2 frame relative to an actual $+0$ frame single-coding region). Also non-coding regions can sometimes give false positives for single-coding – particularly for low $\sum_{\text{tree}} \Lambda$. The latter is partly because, if there are just a few mutations, then they may by chance be consistent with one of the six possible coding frames. However, in all these cases the $\sum_{\text{tree}} \log(\text{LR})$ statistics are typically lower than those for true coding sequences. Hence, if the null model is non-coding, then it is possible to discriminate more strongly against -2 frame and non-coding false positives by taking, for example, $y = x/3$ as the separator rather than $y = 0$. Exactly which separator to use – if other than the natural $y = 0$ – is up to the user to decide, based on the desired balance between false positive and false negative rates. Below, we quote error rates for both $y = 0$ and $y = x/3$ separators. Another approach to achieve more powerful discrimination in these cases is to use the Monte Carlo simulations option on the website. Here, sim-

ulations are used to calculate the expected distributions of the $\sum_{\text{tree}} \log(\text{LR})$ statistic under each of the null and alternative models, and the observed $\sum_{\text{tree}} \log(\text{LR})$ statistic is compared with these two distributions in terms of their respective means and standard deviations.

The success rates for classification are shown in Figure 3 (solid lines: $y = 0$ separator; dashed lines: $y = x/3$ separator). The false positive and false negative rates are summarized, for alignments with only 18–22 mutations, in Table 1. Such an alignment might comprise, for example, five 80 nt sequences with a star-shaped phylogeny and a mean pairwise divergence of 0.1 mutations per nt or, alternatively, a pairwise comparison of two 100 nt sequences with a mean divergence of 0.2 mutations per nt. For such an alignment, the typical false negative rates for identifying single-coding and double-coding regions are 0.00 and 0.15, respectively. The typical false positive rates for predicting single-coding in non-coding regions, predicting double-coding in single-coding regions, and predicting wrong-frame single-coding in single-coding regions are 0.19, 0.05 and 0.03, respectively ($y = 0$ separator; -2 frame excluded). If the $y = x/3$ separator is used whenever the null model is non-coding, then the false negative rate for identifying single-coding rises from 0.00 to 0.04, but the false negative rates for predicting single-coding in non-coding regions and for predicting wrong-frame single-coding in single-coding regions reduce from 0.19 and 0.03 to 0.00 and 0.00, respectively (-2 frame excluded).

Addendum

With regards to the prediction of new genes in ssRNA+ and dsRNA viruses, perhaps the most important statistics are the false positive rates for (a) predicting a single-coding region in a sequence that is actually non-coding, and (b) predicting a +1 or +2 frame overlapping CDS in a sequence that is actually single-coding (in the +0 frame). These statistics are shown for a random sample of the sub-alignments in Figures 4 and 5, respectively. In our search for new virus genes, we prefer to use the selection criteria $\sum_{\text{tree}} \Lambda \geq 20$ and $y \geq x/6$. For these criteria, the probability of falsely predicting a non-overlapping gene is $\leq 4\%$, and the probability of falsely predicting a forward-frame overlapping gene is $\leq 0.7\%$ (Table 1).

References

1. Firth AE, Brown CM: **Detecting overlapping coding sequences with pairwise alignments**. *Bioinformatics* 2005, **21**:282–292.

Figures

Figure 1 - MLOGD statistics for CDSs and non-coding ORFs

The left-hand panel displays the per nt $\sum_{\text{tree}} \log(\text{LR})$ statistic versus $\sum_{\text{tree}} \lambda$. Red points depict data for 37 known CDSs in 14 virus groups (see website for details). Green and blue points depict data for 235 and 169 ORFs, greater than 40 codons in length, that are not known to be coding. The blue points are for CDSs that overlap, at least partially, known CDSs in the -2 read-frame – a frame combination which is liable to result in false positive signals. Apart from these, there is a clear distinction between the known CDSs (positive $\sum_{\text{tree}} \log(\text{LR})$ statistics) and the non-coding ORFs (negative $\sum_{\text{tree}} \log(\text{LR})$ statistics). Even the blue points in general have lower $\sum_{\text{tree}} \log(\text{LR})$ statistics than the red points. The right-hand panel shows the 'summed-over-ORF' rather than 'per nt' statistics. Here, the distinction between the true CDSs and the non-coding ORFs is even clearer, since true CDSs tend to be longer than non-coding ORFs.

Figure 2 - MLOGD sensitivity

Raw $\sum_{\text{tree}} \log(\text{LR})$ (y-axis) versus $\sum_{\text{tree}} \Lambda$ (x-axis) statistics for a small random sample of short sub-regions extracted from known non-coding, single-coding and double-coding regions of virus alignments (see text). The true coding status is indicated by 'actually' (NC = non-coding, SC = single-coding, DC = double-coding). The null and alternative models are similarly indicated. The frames, where shown, are relative to

the frame of the true CDS. In general, points lie above the line when the alternative model agrees with the actual coding status (red points), and otherwise lie below the line (blue points). However, the -2 frame produces many false positives (see text). See [1] or website for frame notation.

Figure 3 - CDS detection success rates
Short sub-regions were extracted from known non-coding, single-coding and double-coding regions of virus alignments (see text) and used to test the success rate of MLOGD for discriminating CDSs from non-coding ORFs as a function of the total number of mutations $\sum_{\text{tree}} \Lambda$ across the sub-alignment. The plot shows the fraction of sub-regions classified as the alternative model for a variety of situations. The true coding status is indicated by 'actually' (NC = non-coding, SC = single-coding, DC = double-coding). The null and alternative models are similarly indicated. The frames, where shown, are relative to the frame of the true CDS. The solid lines are for a $y = 0$ separator and the dashed lines are for a $y = x/3$ separator (see text). The two panels where the alternative model reflects the true-coding status have graphs approaching 1 as $\sum_{\text{tree}} \Lambda$ increases, while most of the other panels – with the exception of the -2 frame (see text) – have graphs approaching 0 as $\sum_{\text{tree}} \Lambda$ increases, as expected. See [1] or website for frame notation.

Figure 4 - False positive rate for predicting a CDS in a sequence that is actually non-coding
Raw $\sum_{\text{tree}} \log(\text{LR})$ (y-axis) versus $\sum_{\text{tree}} \Lambda$ (x-axis) statistics for a random sub-sample of short sub-regions extracted from virus alignments (alignment columns randomized to crudely simulate non-coding sequence; see text). For the selection criteria $\sum_{\text{tree}} \Lambda \geq 20$ and $y \geq x/6$ (red lines) the probability of falsely predicting a CDS in non-coding sequence is $\leq 4\%$. (NC = non-coding, SC = single-coding.)

Figure 5 - False positive rate for predicting a +1 or +2 frame overlapping CDS in a sequence that is actually single-coding in the +0 frame
Raw $\sum_{\text{tree}} \log(\text{LR})$ (y-axis) versus $\sum_{\text{tree}} \Lambda$ (x-axis) statistics for a random sub-sample of short sub-regions extracted from single-coding virus alignments (see text). For the selection criteria $\sum_{\text{tree}} \Lambda \geq 20$ and $y \geq x/6$ (red lines) the probability of falsely predicting a $+1$ or $+2$ frame overlapping CDS in a sequence that is actually single-coding in the $+0$ frame is $\leq 0.7\%$. (SC = single-coding, DC = double-coding.)

Tables

Table 1 - CDS detection error rates
Table showing the error rates for CDS detection with MLOGD, as estimated from real sequence data. The first column gives the true coding status (NC = non-coding, SC = single-coding, DC = double-coding). The second and third columns give the null and alternative models, respectively. The frames, where shown, are relative to the frame of the true CDS. The error rate is the fraction of alignments mis-classified. The alignments are of varying lengths (≥ 10 codons) but all have between 18 and 22 mutations in total across the alignment. Figures in bold are those alignments where the alternative model is the correct model (i.e. false negative rates). The other figures are false positive rates. See text for details on the $y = 0$, $y = x/6$ and $y = x/3$ separators.

Actual coding	Null model	Alternative model (frame)	Error rate		
			$y = 0$	$y = x/3$	$y = x/6$
NC	NC	SC	0.19	0.00	0.041
SC	NC	SC (+0)	**0.00**	**0.04**	
SC	NC	SC (+1)	0.02	0.00	
SC	NC	SC (+2)	0.03	0.00	
SC	NC	SC (−0)	0.04	0.00	
SC	NC	SC (−1)	0.01	0.00	
SC	NC	SC (−2)	0.96	0.48	
DC	SC	DC	**0.15**		
SC	SC	DC (+1)	0.04		0.0065
SC	SC	DC (+2)	0.05		0.0064
SC	SC	DC (−0)	0.08		
SC	SC	DC (−1)	0.04		
SC	SC	DC (−2)	0.75		

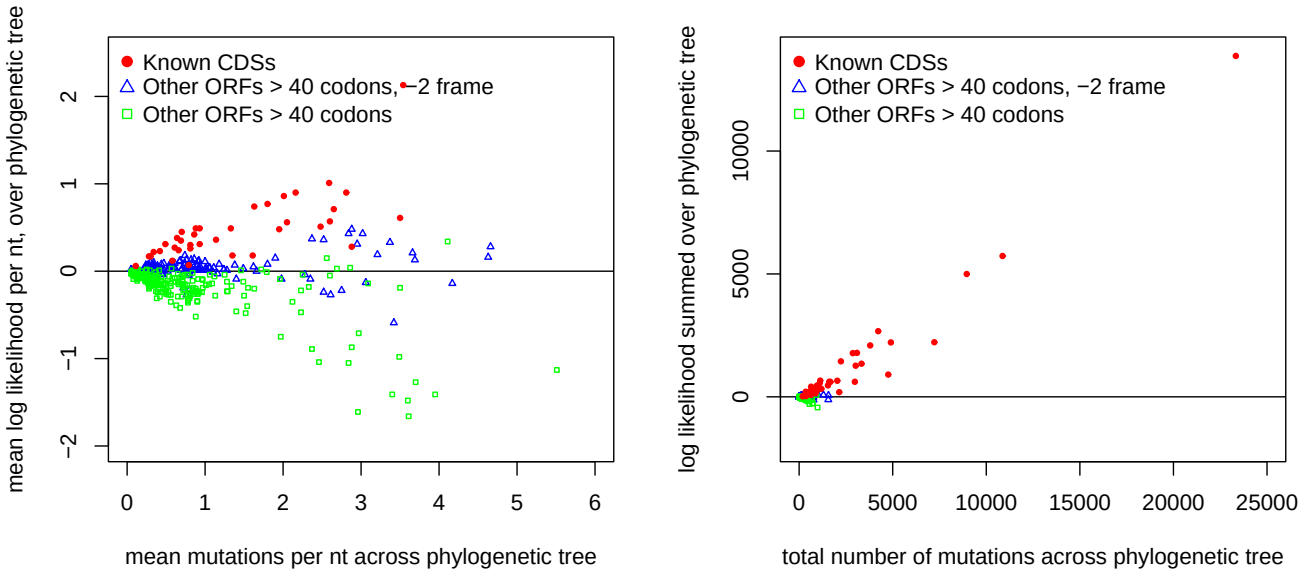

Figure 1:

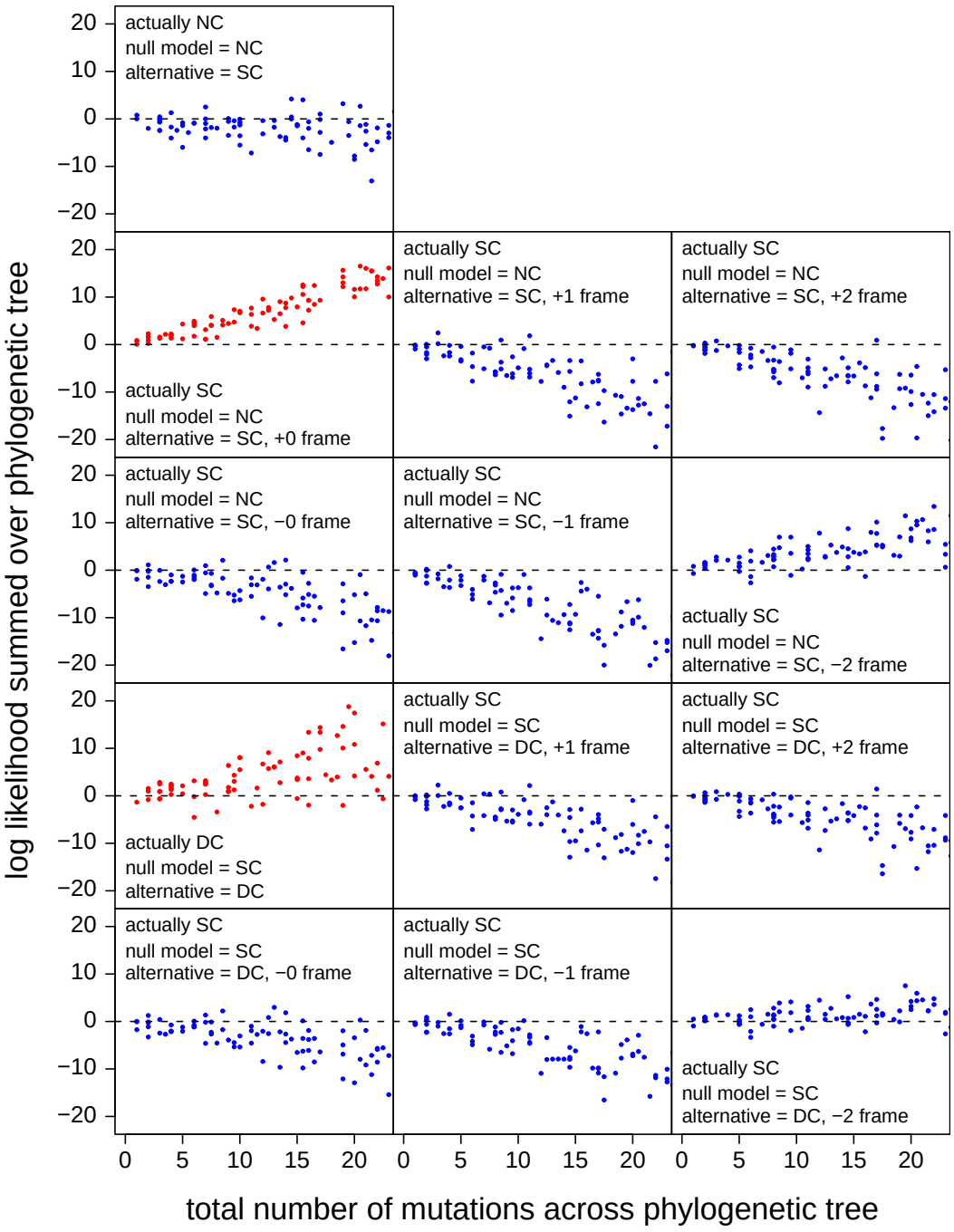
log likelihood summed over phylogenetic tree
total number of mutations across phylogenetic tree
actually NC
null model = NC
alternative = SC
actually SC
null model = NC
alternative = SC, +1 frame
actually SC
null model = NC
alternative = SC, +2 frame
actually SC
null model = NC
alternative = SC, +0 frame
actually SC
null model = NC
alternative = SC, −0 frame
actually SC
null model = NC
alternative = SC, −1 frame
actually SC
null model = NC
alternative = SC, −2 frame
actually DC
null model = SC
alternative = DC
actually SC
null model = SC
alternative = DC, +1 frame
actually SC
null model = SC
alternative = DC, +2 frame
actually SC
null model = SC
alternative = DC, −0 frame
actually SC
null model = SC
alternative = DC, −1 frame
actually SC
null model = SC
alternative = DC, −2 frame

Figure 2:

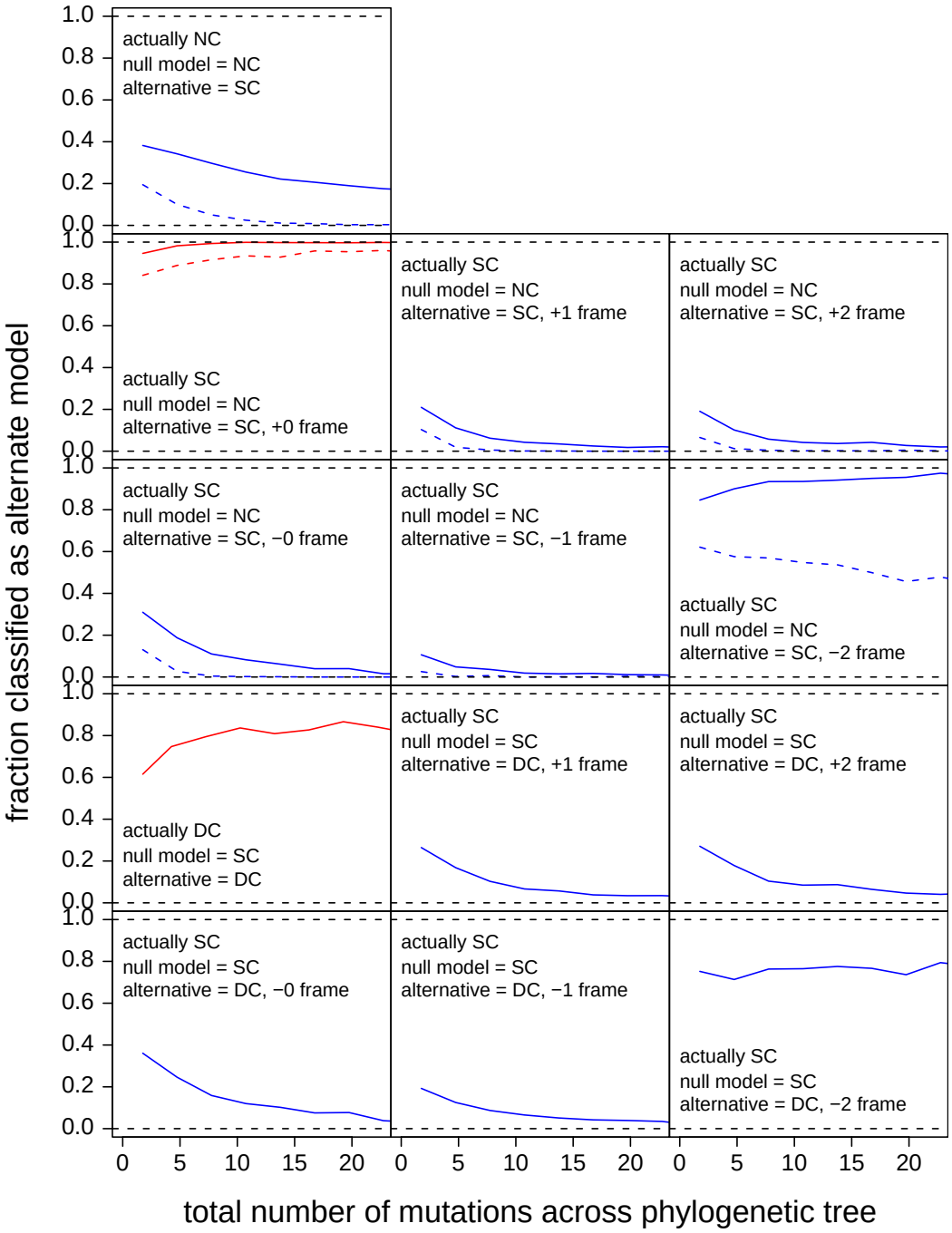
fraction classified as alternate model
total number of mutations across phylogenetic tree
actually NC
null model = NC
alternative = SC
actually SC
null model = NC
alternative = SC, +0 frame
actually SC
null model = NC
alternative = SC, +1 frame
actually SC
null model = NC
alternative = SC, +2 frame
actually SC
null model = NC
alternative = SC, −0 frame
actually SC
null model = NC
alternative = SC, −1 frame
actually SC
null model = NC
alternative = SC, −2 frame
actually DC
null model = SC
alternative = DC
actually SC
null model = SC
alternative = DC, +1 frame
actually SC
null model = SC
alternative = DC, +2 frame
actually SC
null model = SC
alternative = DC, −0 frame
actually SC
null model = SC
alternative = DC, −1 frame
actually SC
null model = SC
alternative = DC, −2 frame

Figure 3:

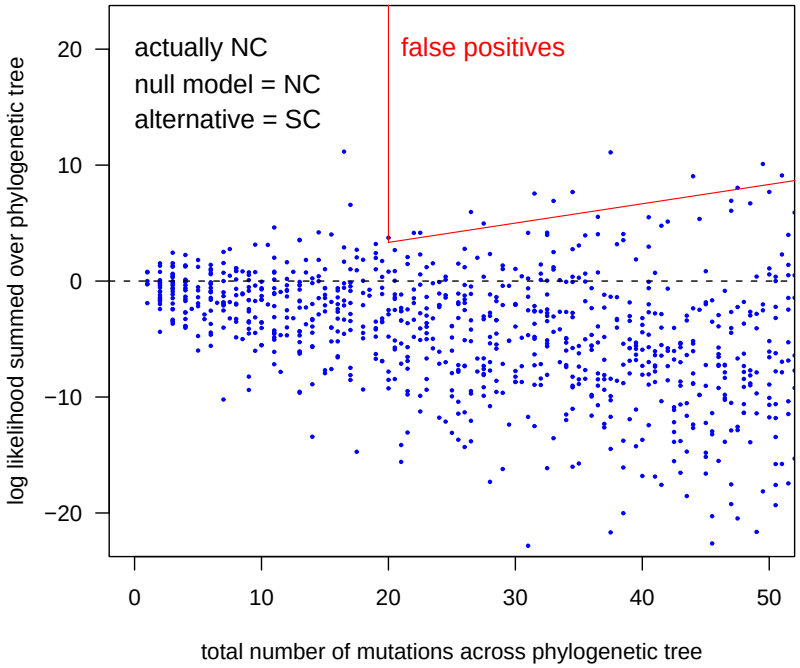

Figure 4:

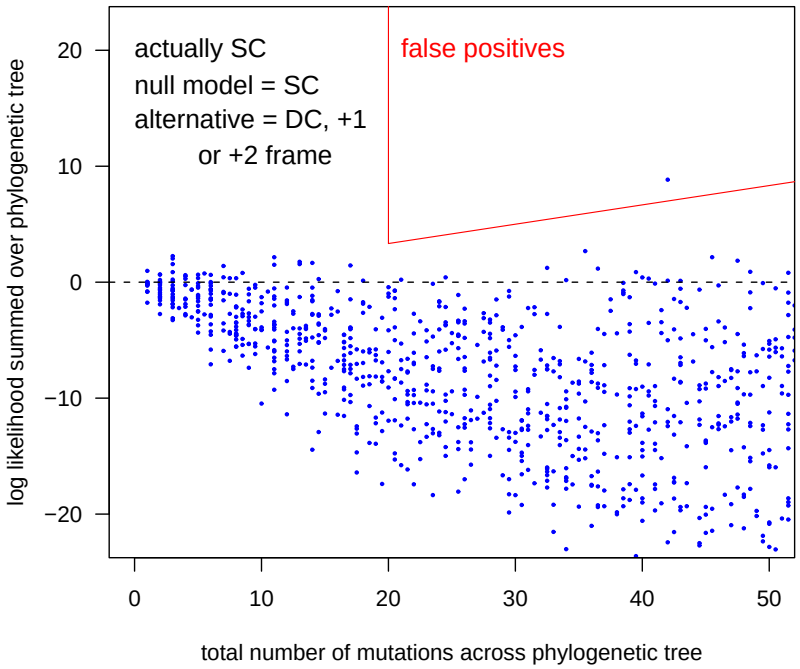

Figure 5: