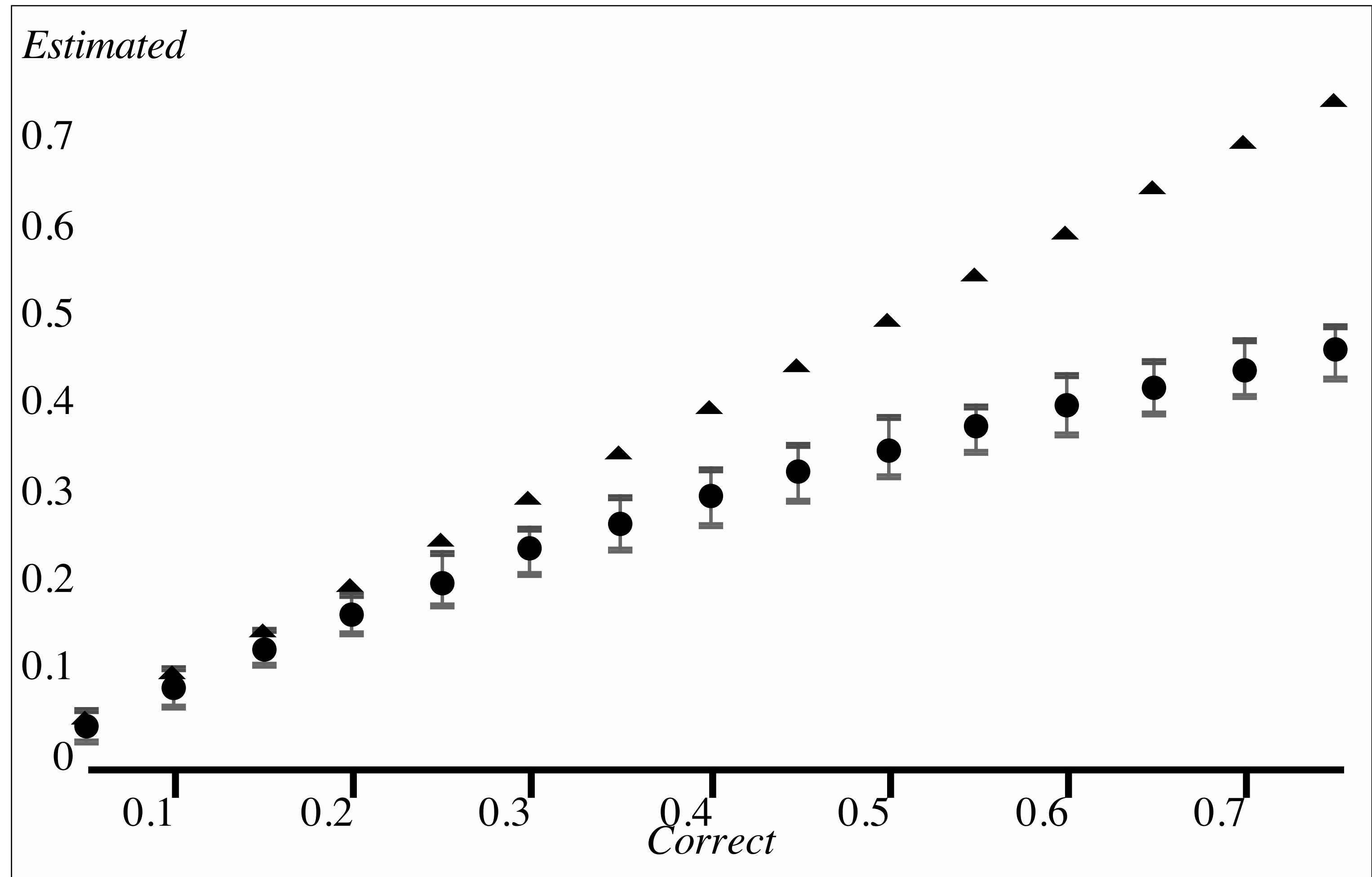


# NG86 limitations: underestimation of substitution counts for higher divergence levels

Simulated divergence vs that estimated by p-distance

- Simulated 100 replicates of 1000 nucleotide long sequences for various divergence levels (substitutions/site)
- Even for divergence of 0.25 (1/4 sites have mutation on average), p-distance already underestimates the true level: 0.2125 (0.19–0.241 95% range)
- Underestimation becomes progressively worse for larger divergence levels



# NG86 limitations: ignoring phylogenies

†



Fig. 1.1. Effect of phylogeny on estimating synonymous and nonsynonymous substitution counts in a dataset of Influenza A/H5N1 haemagglutinin sequences. Using the maximum likelihood tree on the left, the observed variation can be parsimoniously explained with one nonsynonymous substitution along the darker branch, whereas the star tree on the right involves at least two.