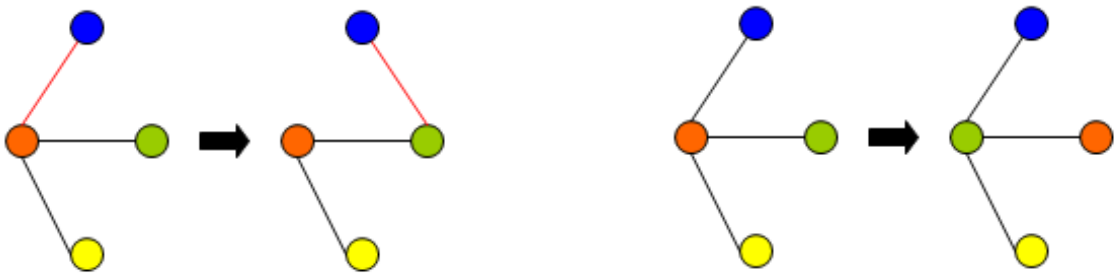
Supplementary Material for
"Evolutionary analysis reveals low coverage as the major challenge for protein interaction network alignment"

Waqar Ali, Charlotte M. Deane

This document contains the supplementary figures along with detailed legends/explanations for the paper "Modeling the effect of interactome evolution on network alignment". Figures 1-6 relate to clarifications and additional analysis for the Human-Yeast pair of networks discussed in the main text while Figures 7 and 8 illustrate the applicability of our methods on Human-Fly and Fly-Yeast pairs.



a **b**

Fig1. Uniform Error Models: a) Edge rewiring error and (b) Node relabeling error. The edge rewiring error removes one true interaction (blue-orange) and inserts one spurious interaction (blue-green) at each step. The node relabeling error swaps the identities of two proteins in the network at each step. This can affect multiple edges simultaneously. Swapping green with orange removes two true interactions (blue-orange, yellow-orange) and inserts two spurious interactions (blue-green, yellow-green).

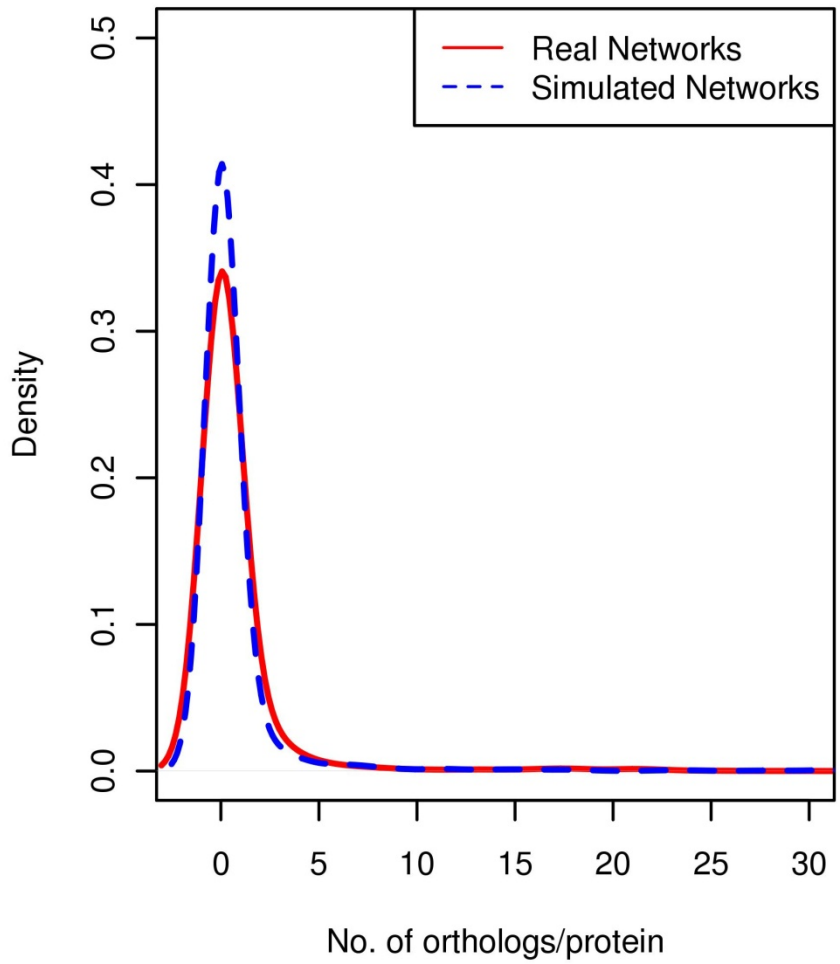


Fig 2 Real vs. simulated ortholog distribution: Distribution of orthologs between the real yeast and human networks (bidirectional blast E-value < 10^{-10}), and simulated yeast and human networks using our orthology evolution mechanism.

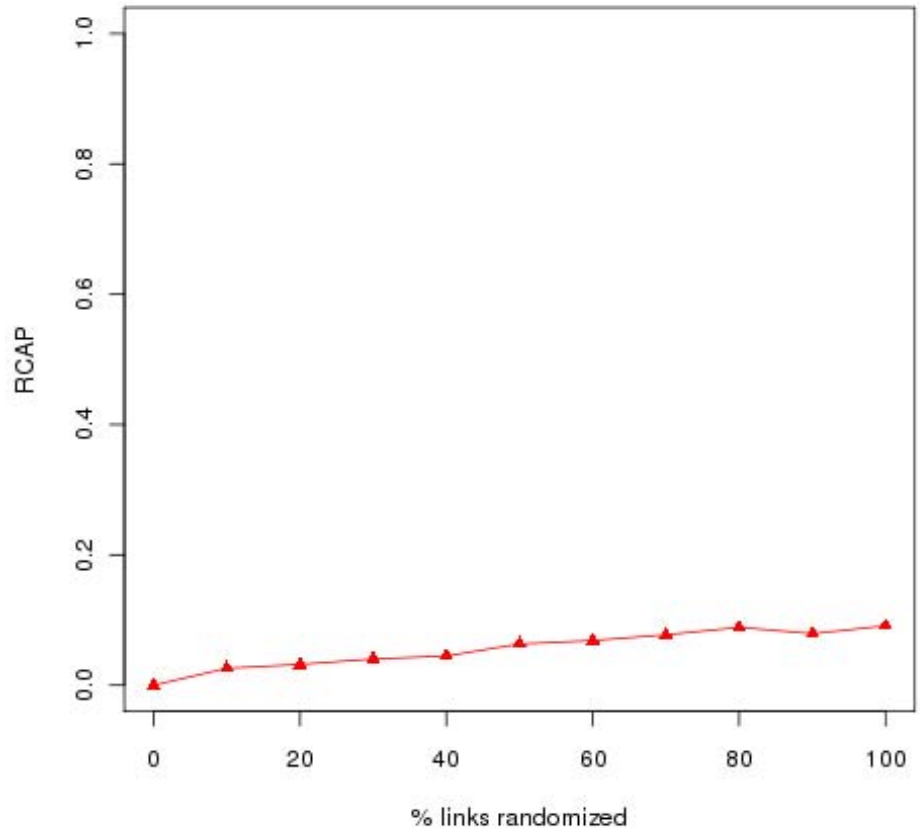
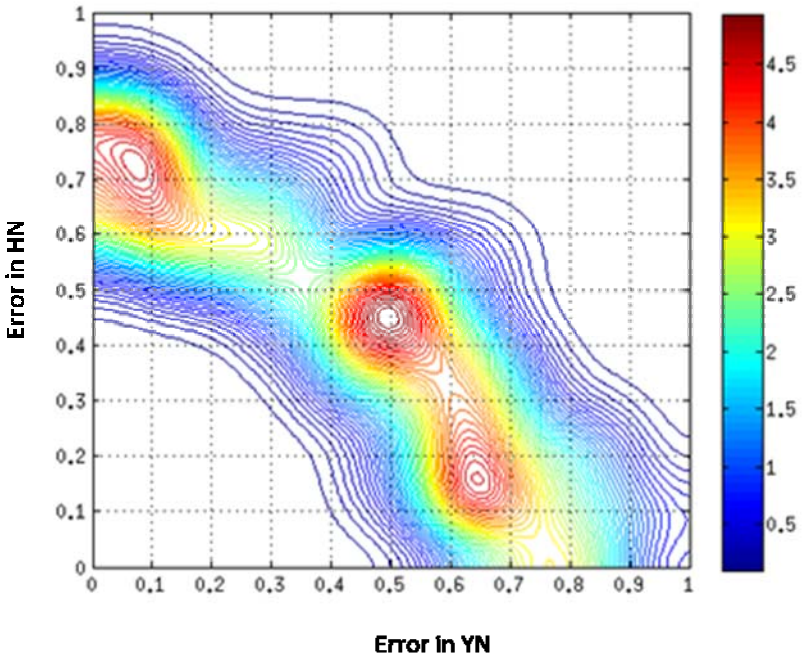
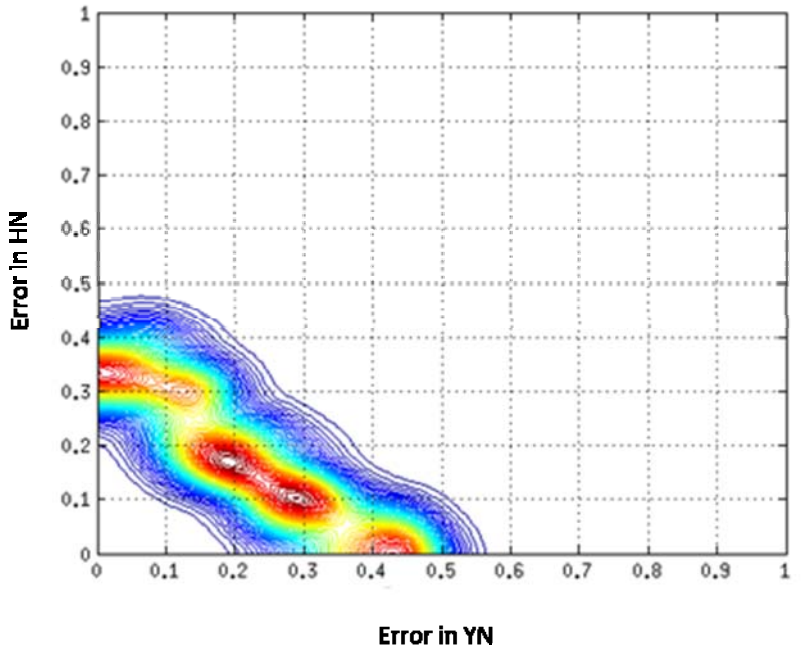


Fig 3. Effect of ancestral topology on alignment: The change in alignment statistics due to random shuffling of the ancestral network topology. Let n_0 be the original ancestral network as defined in the paper and n_x be the same ancestral network with x% of its links randomly rewired. Also, let An_0 and An_x be the alignments of pairs of networks evolved from these ancestors and Ar be the alignment of a pair of experimental networks (human and yeast in this case). The RCAP (Relative Change in Alignment Properties) is then defined as dist(An_0, An_x)/dist(An_0, Ar) where the distance measure, dist, between the alignments is the one described in the main text. RCAP essentially measures the effect of changes in the ancestral network topology on the alignment of its descendants, normalized by the distance between real and simulated alignments. As seen in the figure, even for 100% randomization, the change in alignment statistics is very small relative to the distance between real and simulated alignments. The effect of ancestral network topology on our error estimates is thus negligible.



a



b

Fig 4. Error estimates for human and yeast networks using NWBlast (a) without sampling and (b) with sampling: The error density curves for human and yeast networks when alignment is done using NWBlast instead of MaWISh. The curves indicate roughly similar values as results in the main text although the density estimation is not as sharp. This is due to the much smaller number of samples used in density estimation, as NWBlast was found to be far slower than MaWISh.

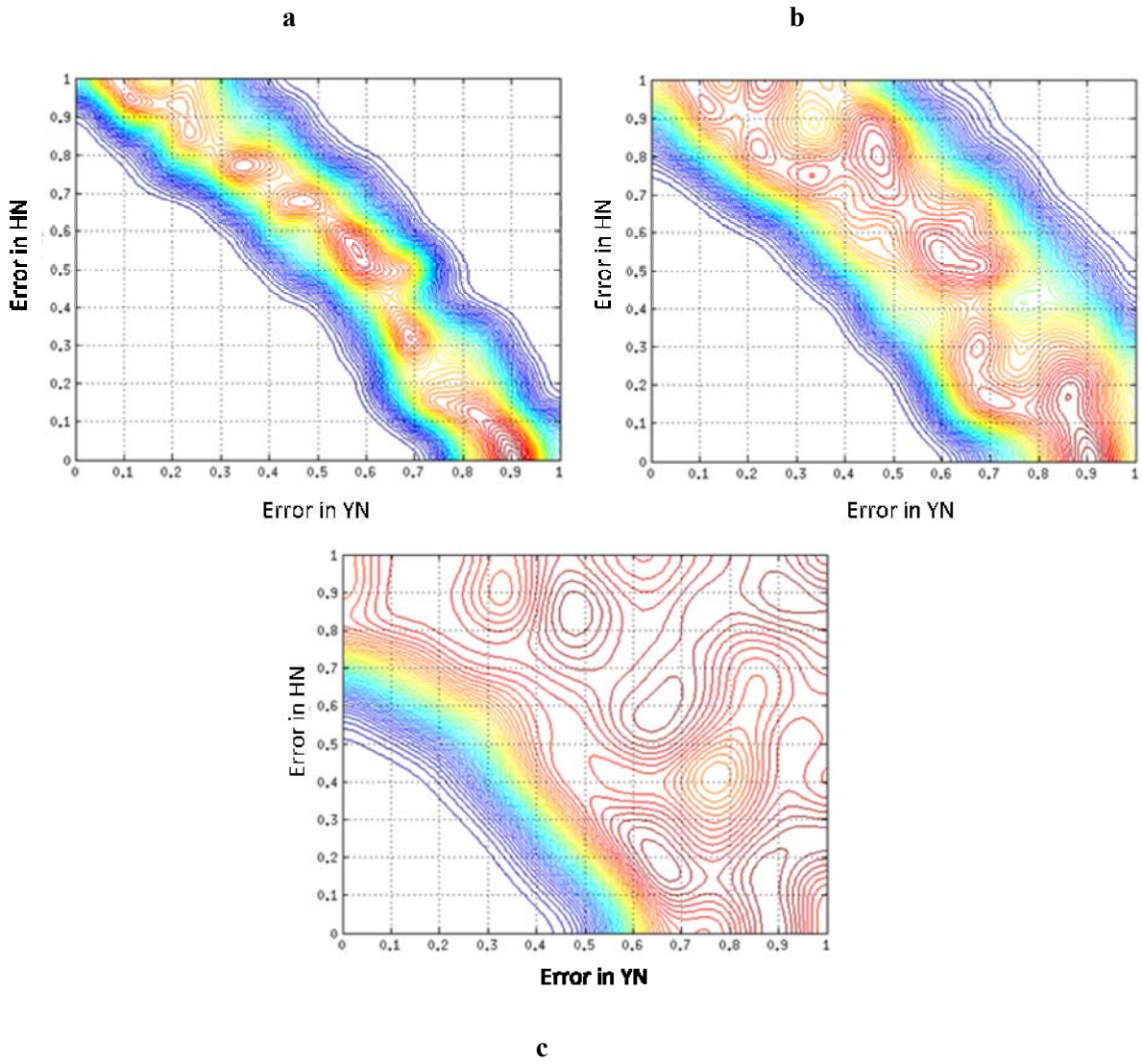


Fig 5. Effect of threshold δ on error density estimation: This figure illustrates how using higher values of the threshold δ in Algorithm 1 negatively affects the error density estimates. Plots a-c show error estimates with the threshold set to 0.005, 0.01 and and 0.02 respectively. Note that lower threshold values require that a much higher number of samples be generated, as the proportion of samples accepted is very small. For instance the density estimates in the main text were calculated by generating 5000 random samples from the two dimensional error parameter space and accepting only those with distance less than 0.001 to real alignments.

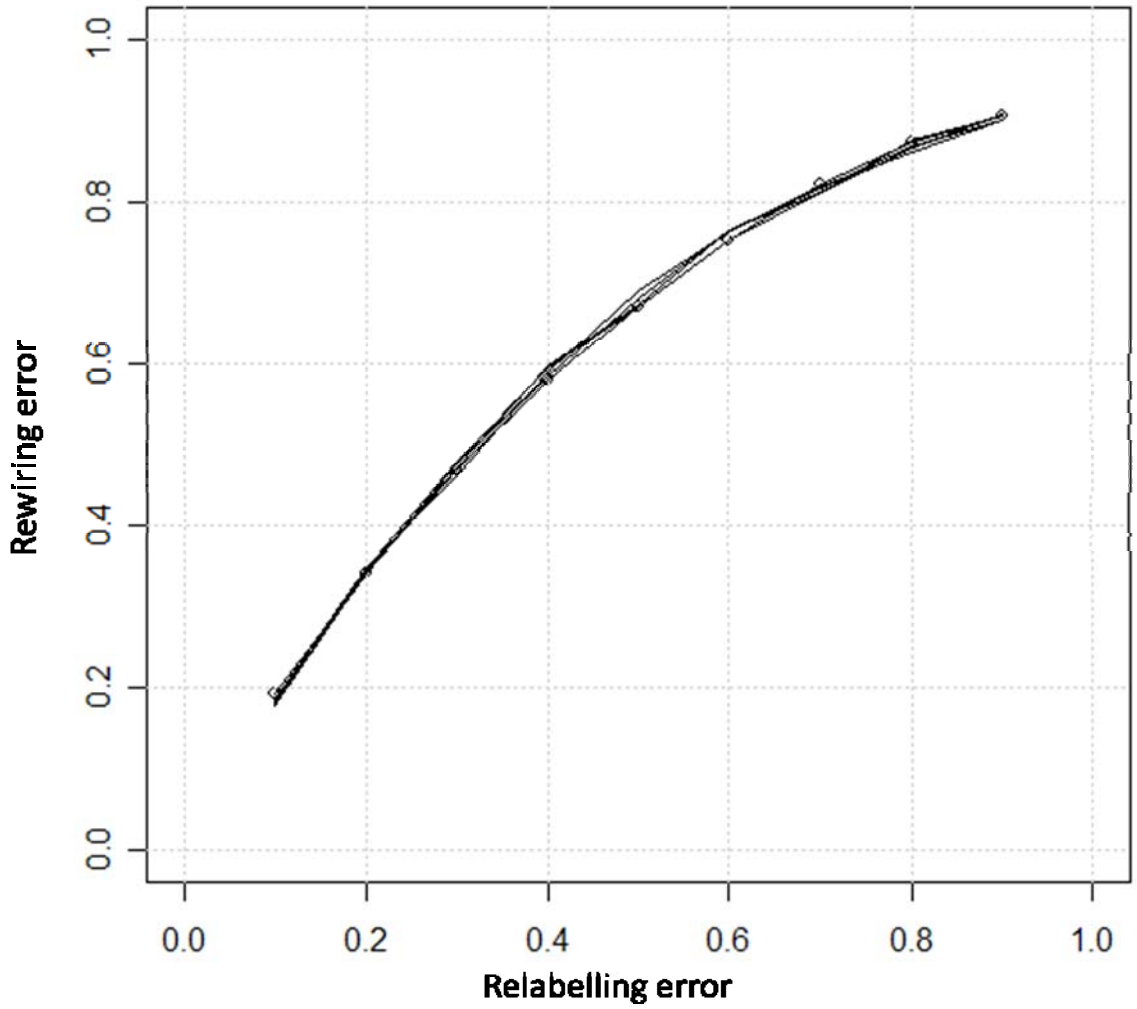
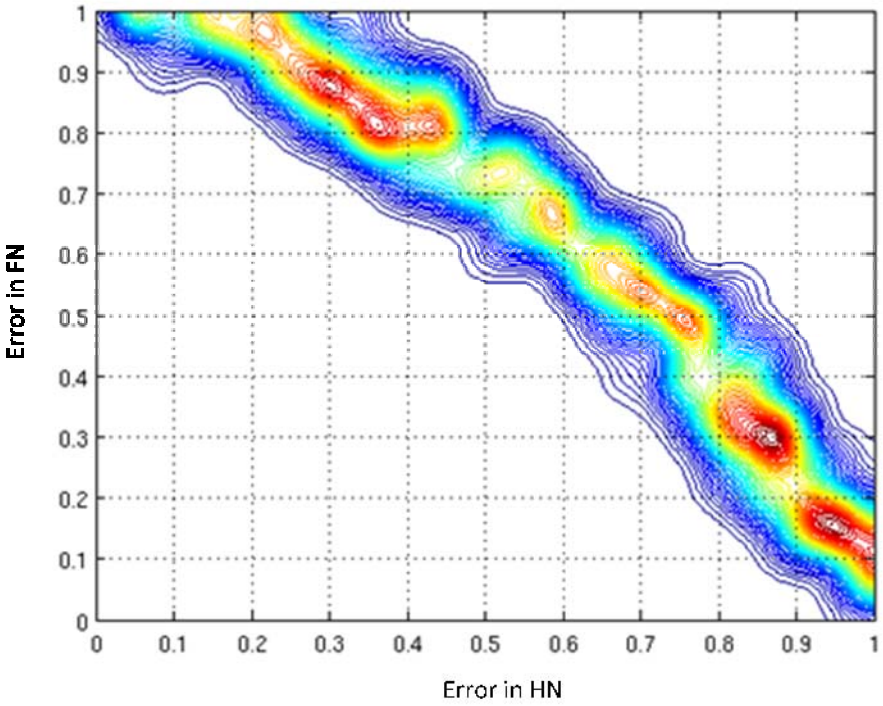
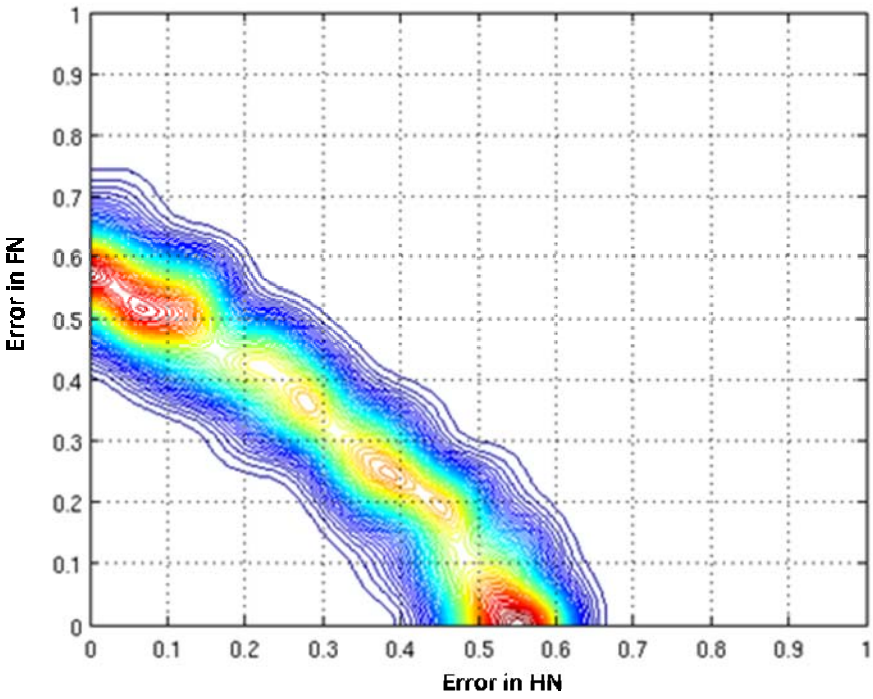


Fig 6. Relationship between relabeling and rewiring error: Node relabeling can affect multiple edges at each step. This leads to a non-linear relationship between the two types of error.

a

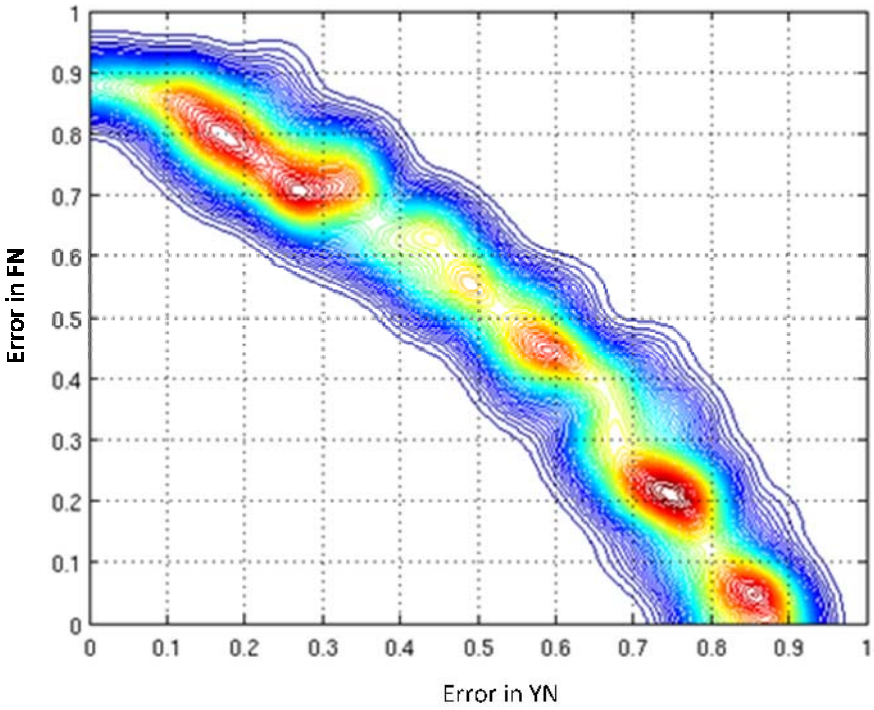


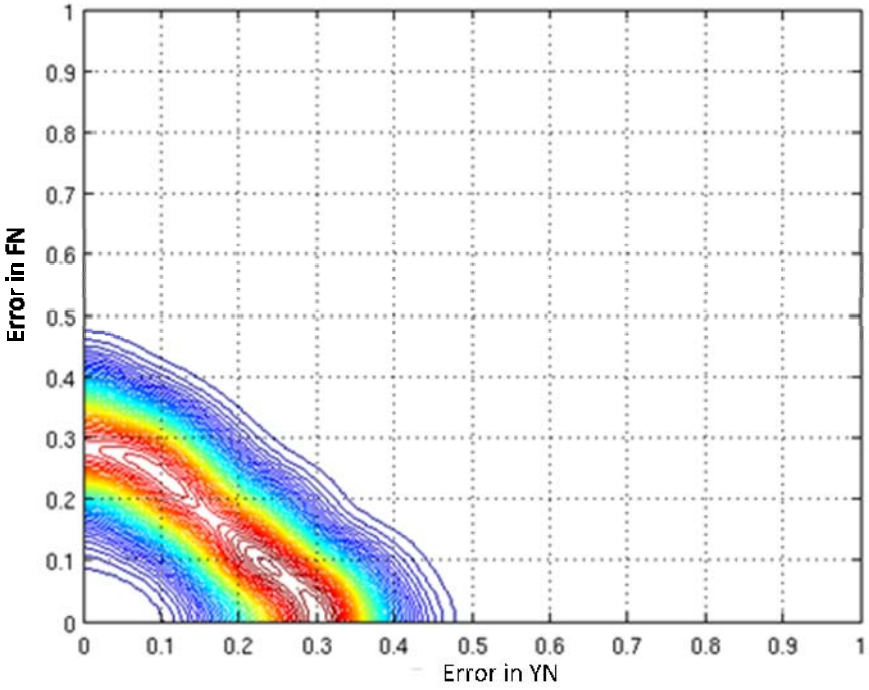


b

Fig 7. Error estimates for Human-Fly using MaWISh (a) without sampling and (b) with sampling: Error estimates for Human and Fly networks using the methods in the main text. The error bands in this case span higher values compared to the Human-Yeast case. This is probably because current experimental data-sets for both these species are substantially incomplete compared to Yeast which is believed to be reasonably complete.

a





b

Fig 8. Error estimates for Yeast-Fly using MaWISh (a) without sampling and (b) with sampling: These plots show similar values as the Human-Yeast analysis in the main text. Observed in context with the Human-Fly plots, it appears that the Fly and Human networks share similar levels of incompleteness while the Yeast data-sets suffer less from false negatives.