Sweeps in time: leveraging the joint distribution of branch lengths

Supplementary Figures

Gertjan Bisschop[1], Konrad Lohse[1], and Derek Setter*[1]

[1]Institute of Evolutionary Biology, University of Edinburgh, UK

* correspondence to derek.setter@ed.ac.uk

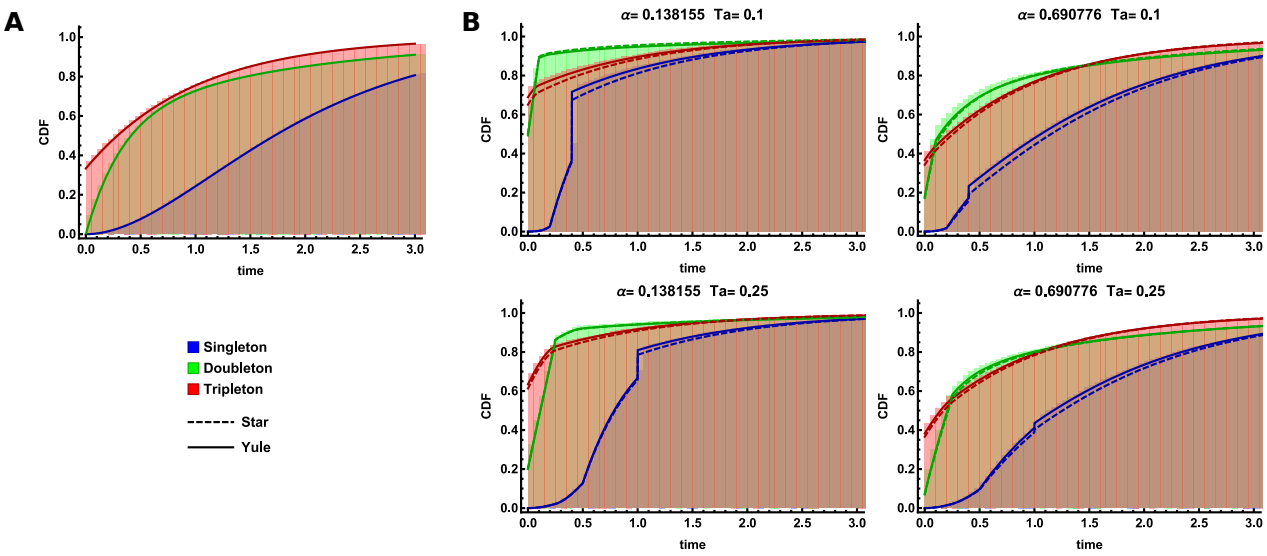


Figure S1: **CDF for marginal branch lengths for** $n = 4$. Companion to Figure 4 of the main text.

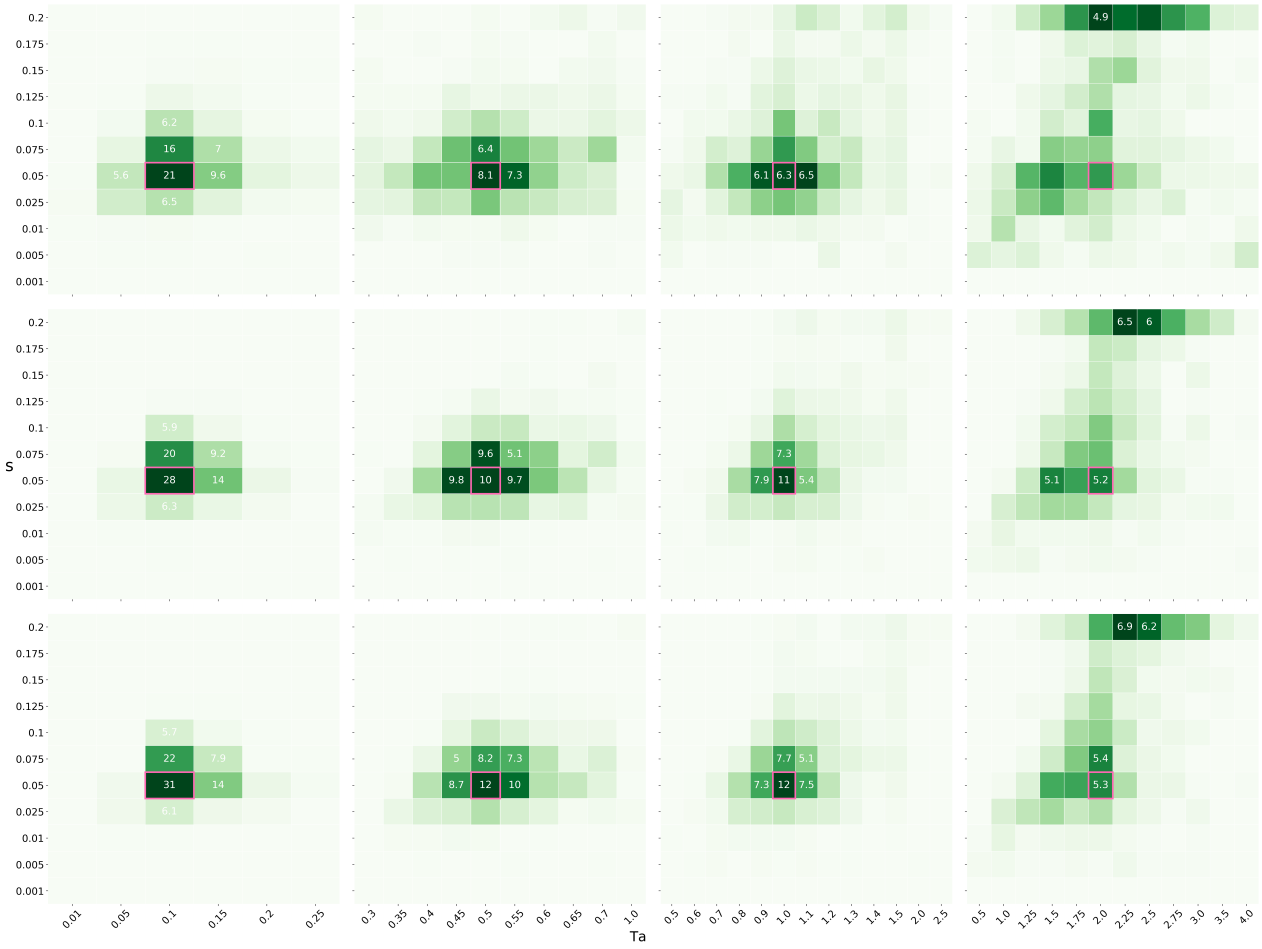


Figure S2: **The effect of sample size on the accuracy to infer sweep parameters, strong selection $s = 0.05$, star-like approximation.** The top row shows the results for $n = 4$ (from Figure 7, panel A), the middle row for $n = 12$, and the bottom for $n = 20$. From left to right, the panels represent increasing sweep age ($T_a = 0.1, 0.5, 1.0, 2.0$). Numbers show the percentage of replicates ($> 4.5\%$) associated with a particular parameter combination. The true parameter combination is indicated by a pink square.

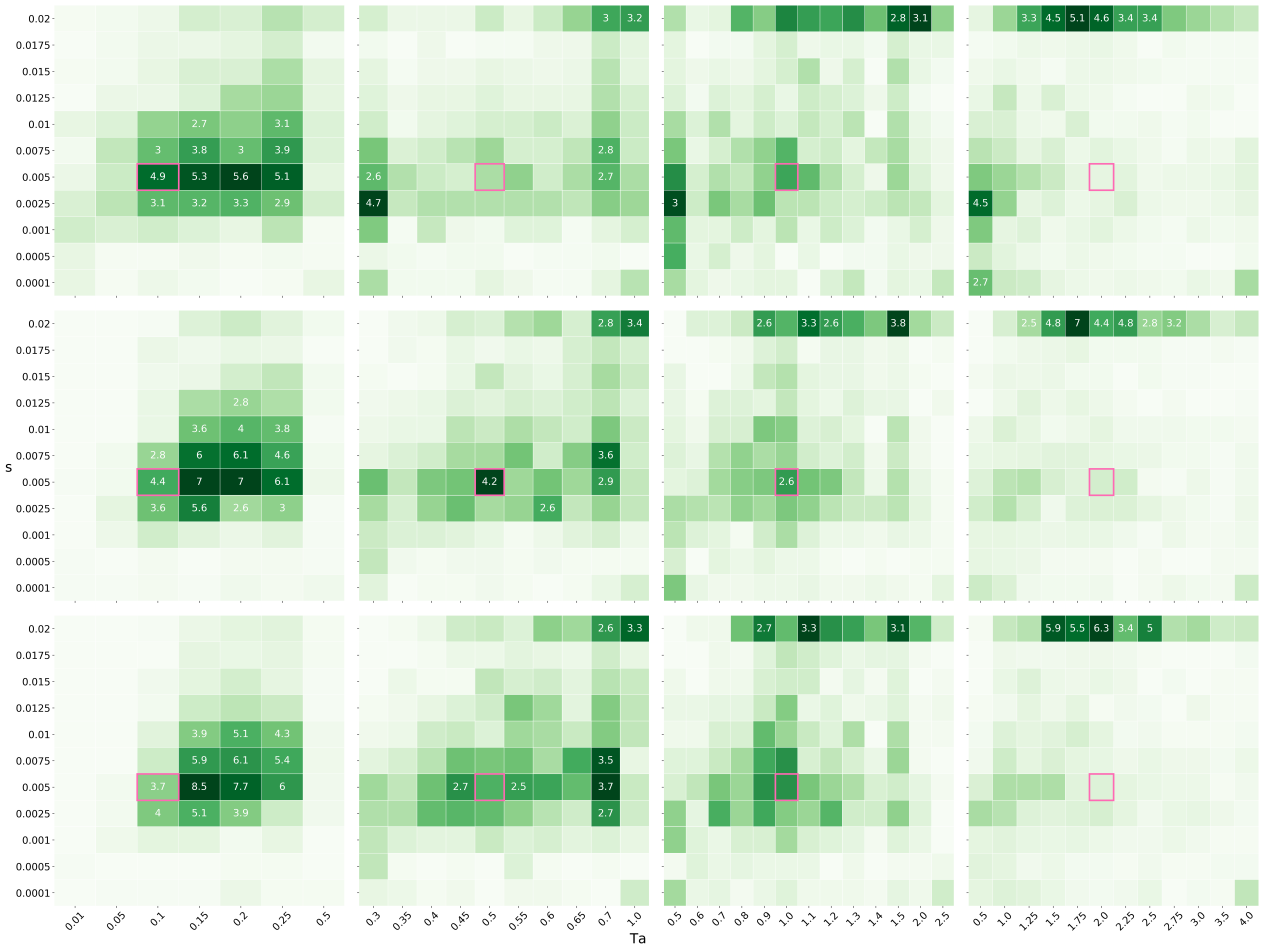


Figure S3: **The effect of sample size on the accuracy to infer sweep parameters, weak selection $s = 0.005$, star-like approximation.** The top row corresponds to Figure 7 panel B. Sample size increases from top to bottom; the age of the sweep increases from left to right. See Figure S2 for a full description.

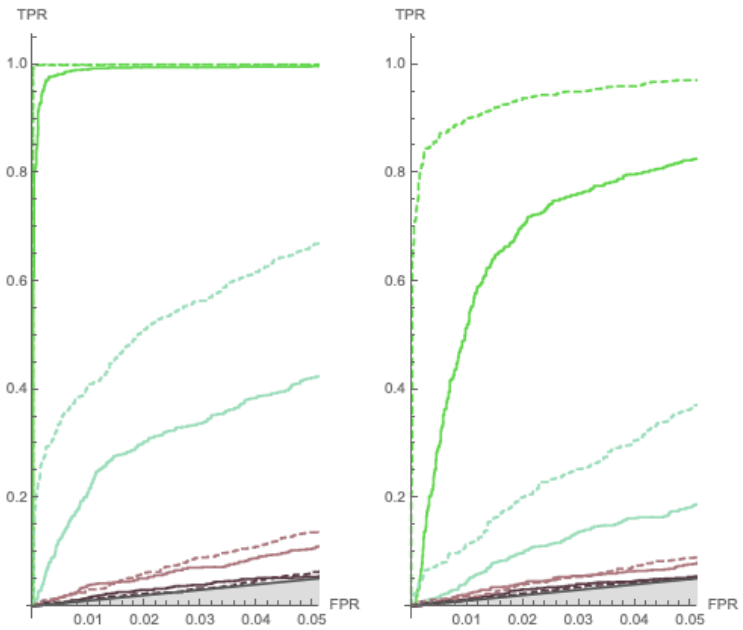


Figure S4: **ROC curve, model with $T_a = 0$, star-like approximation**: Misspecifying a model of a recent sweep ($T_a = 0$) reduces the power to detect old sweeps. Figure 6 shows analogous results for the same set of simulations when fitting the correct model, i.e. including T_a in the inference. As before, colors indicate the age of the simulated sweep: $Ta = 0.1$ (green), 0.5 (lighter green), 1.0(light brown), 2.0 (dark brown). Left: strong selection ($s = 0.05$), right: weak selection ($s = 0.005$), sample size $n = 4$ (full line), and 12 (dashed line).

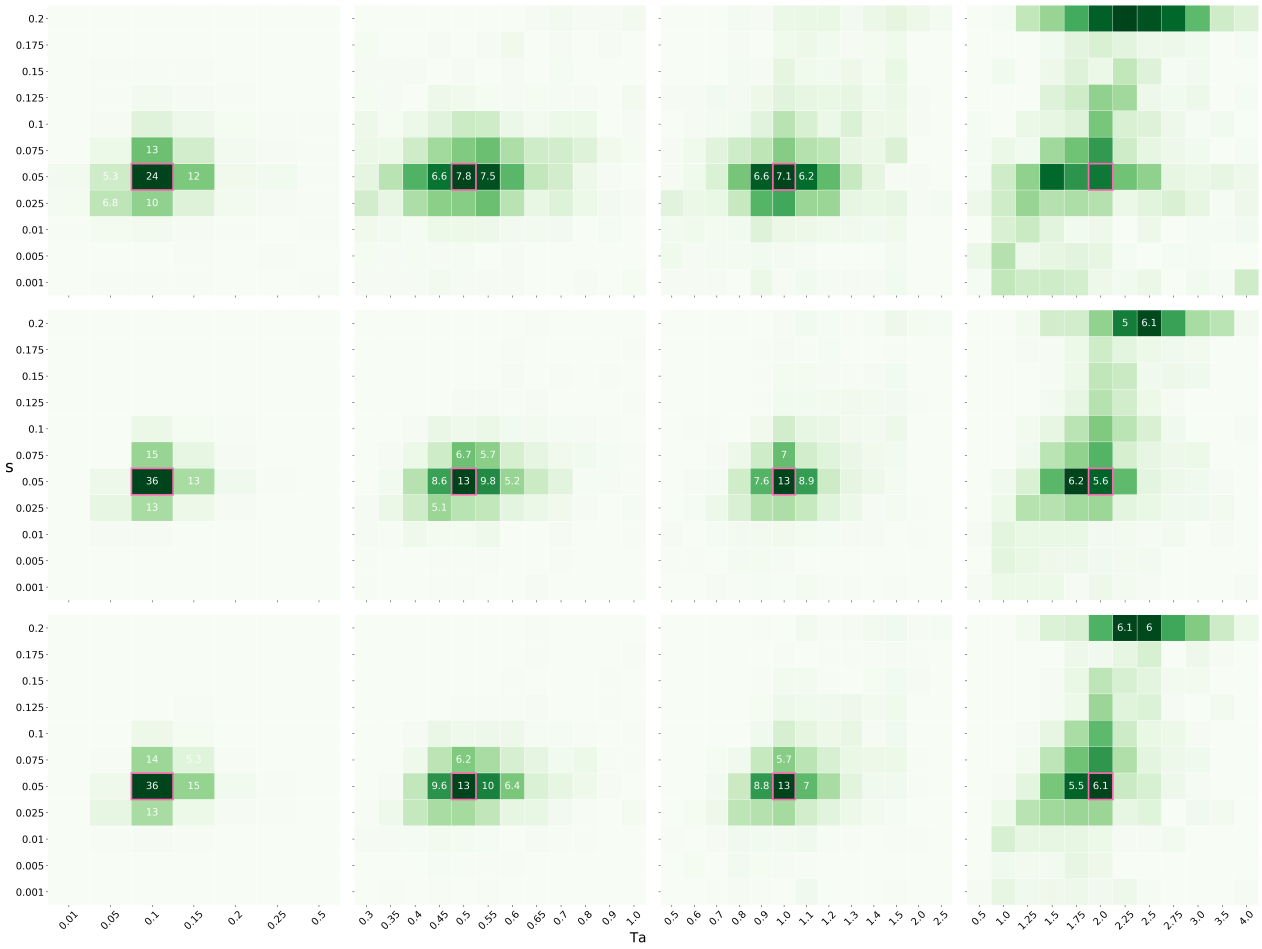


Figure S5: **Heatmap instantaneous Yule approximation, strong selection** $s = 0.05$**.** Results of the gridded optimization. Panels represent different sweep ages ($Ta = 0.1, 0.5, 1.0, 2.0$ from left to right). Rows from top to bottom show results for different sub-sample sizes ($n = 4, 12, 20$). Numbers show the percentage of replicates ($> 4.5\%$) associated with a particular parameter combination. The true parameter combination is indicated by a pink square.

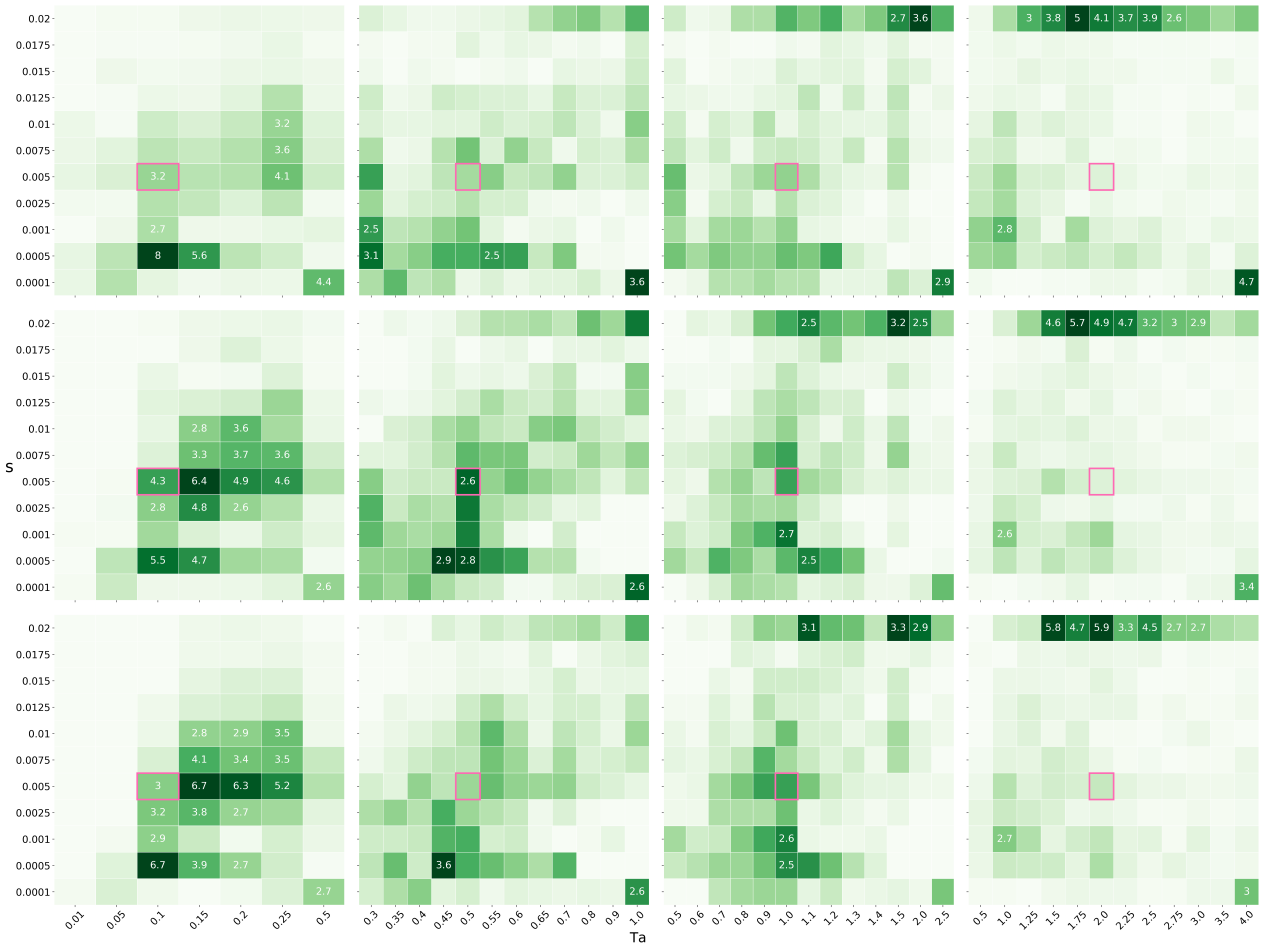


Figure S6: **Heatmap instantaneous Yule approximation, weak selection** $s = 0.005$**.** Results of the gridded optimization. Panels represent different sweep ages ($Ta = 0.1, 0.5, 1.0, 2.0$ from left to right). Rows from top to bottom show results for different sub-sample sizes ($n = 4, 12, 20$). Numbers show the percentage of replicates ($> 4.5\%$) associated with a particular parameter combination. The true parameter combination is indicated by a pink square.

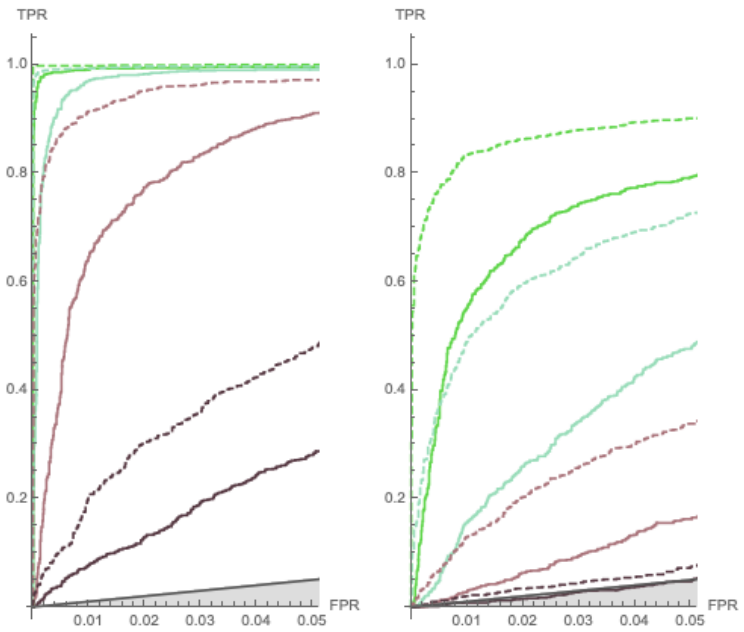


Figure S7: **ROC, instantaneous Yule approximation**: Plotting the rate of true positives against the rate of false negatives ($Ta = 0.1$ (green), 0.5 (lighter green), 1.0 (light brown), 2.0 (dark brown)), the strength of selection (left $s = 0.05$, right $s = 0.005$) and sample size $n = 4$ (full line), 12 (dashed).

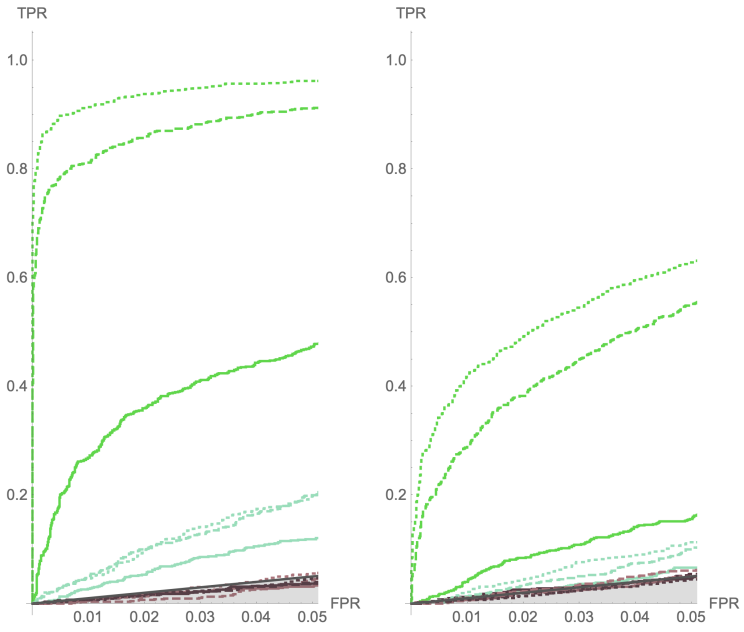


Figure S8: **ROC, SweepFinder2**: Plotting the rate of true positives against the rate of false negatives after applying SweepFinder2 to the same set of simulations used for Figure 6 ($Ta = 0.1$ (green), 0.5 (lighter green), 1.0 (light brown), 2.0 (dark brown)), the strength of selection (left $s = 0.05$, right $s = 0.005$) and sample size $n = 4$ (full line), 12 (dashed).

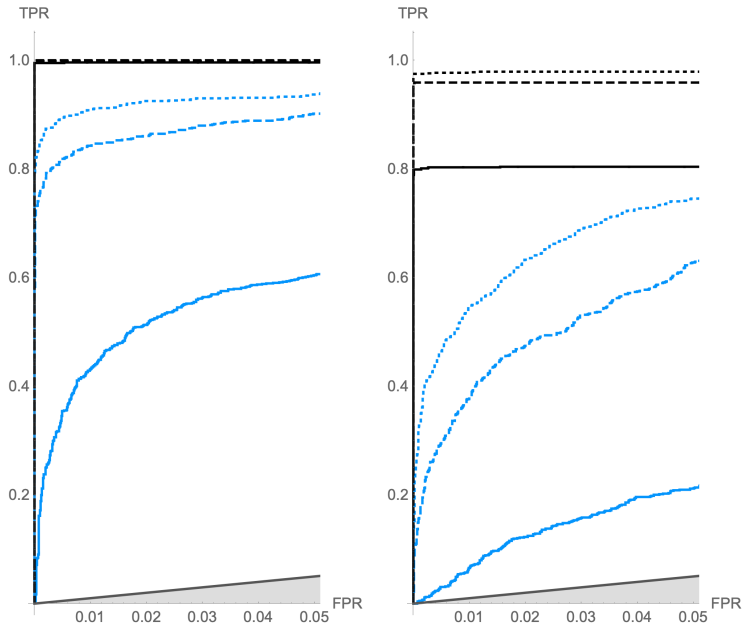
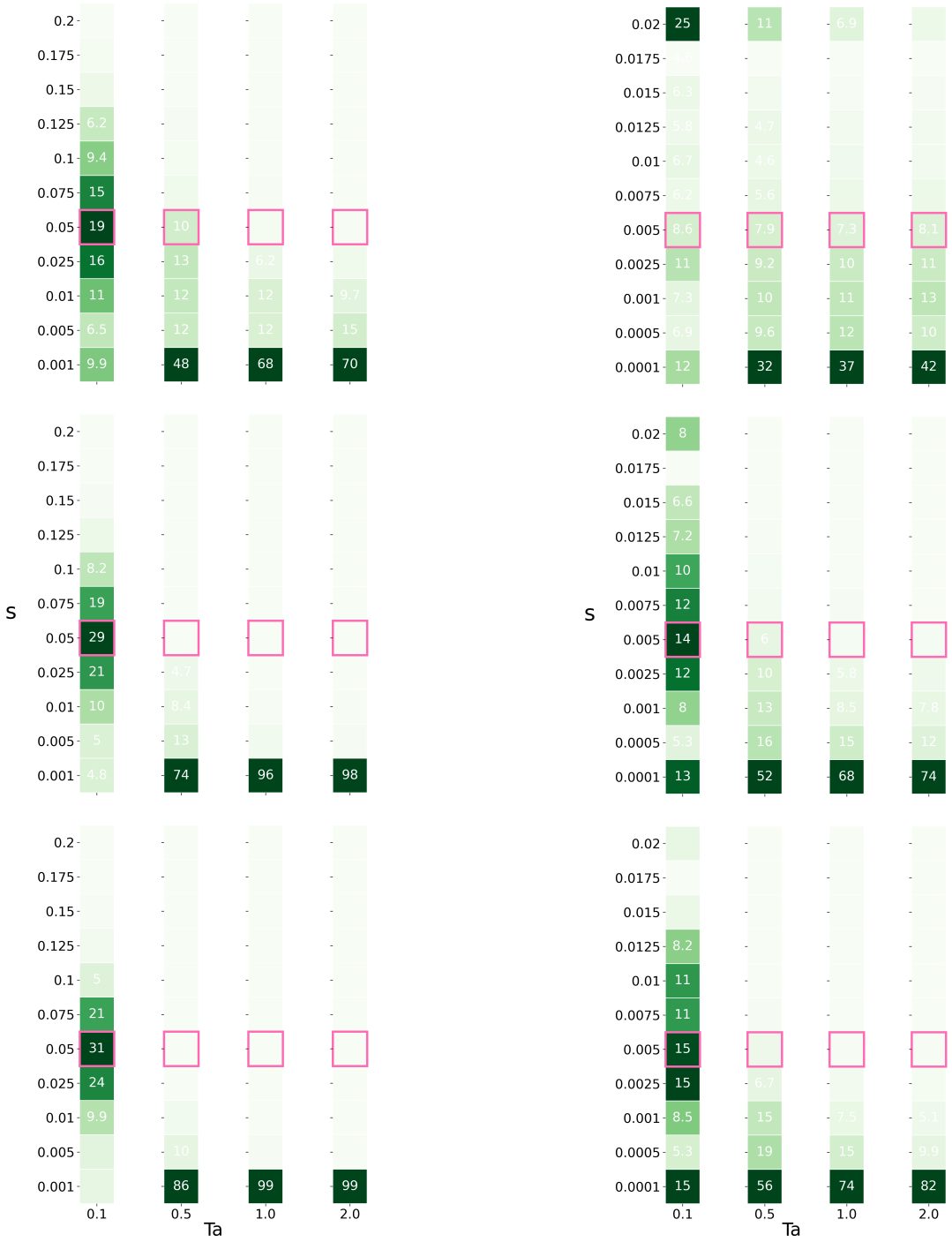


Figure S9: **ROC, sims** $T_a = 0$: Inference based on the bSFS (black) has greater power than SweepFinder2 (blue) which uses the unfolded SFS to distinguish sweep regions from neutral regions. Data were simulated under the model of a recent sweep $T_a = 0$ as assumed by SweepFinder2 for different sample sizes: $n = 4$ (full line), 12 (dashed line) or 20 (dotted line).



(a) strong selection, s=0.05

(b) weak selection, s=0.005

Figure S10: **Heatmaps SweepFinder2**: Estimates for the strength of selection by SweepFinder2. The true, simulated parameter combination is indicated by the pink squares. The simulations analysed are the same that were used to assess the power to infer sweep parameters from the bSFS (Figure S2 and S3).

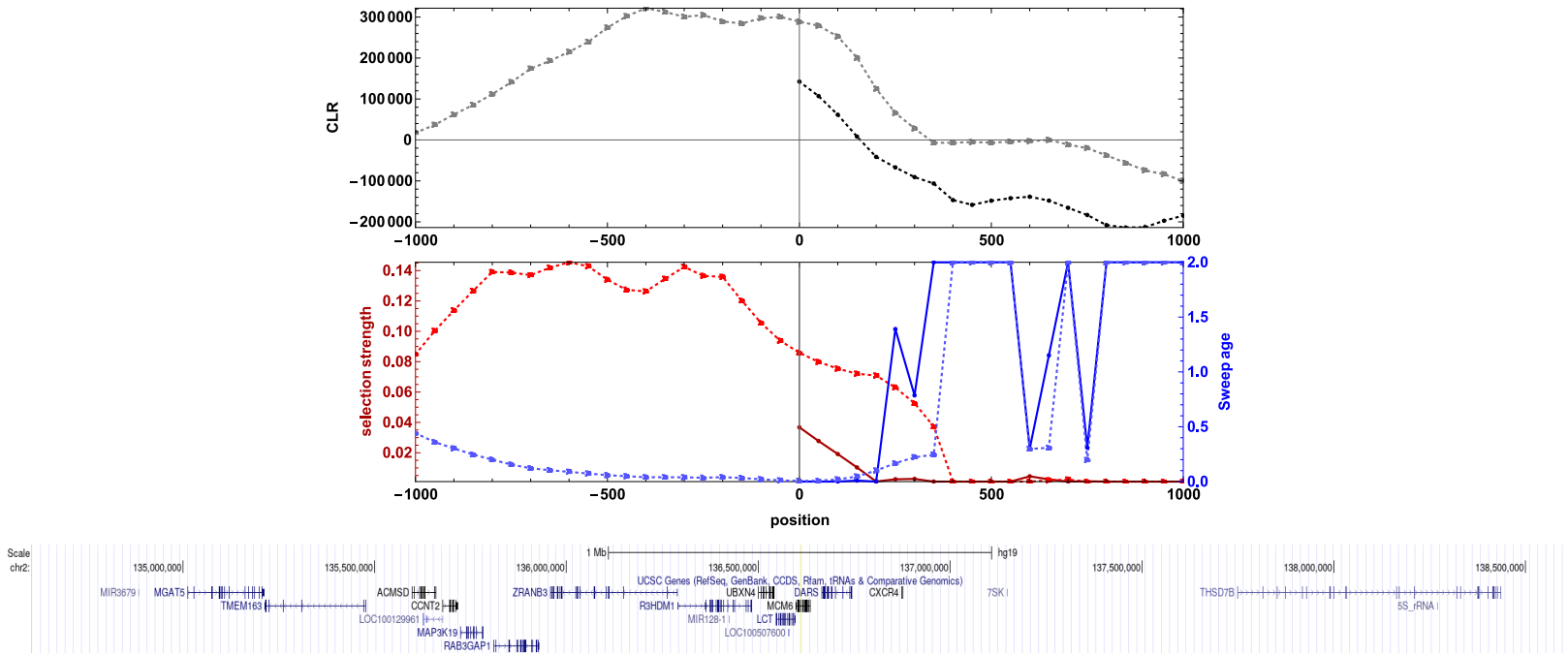


Figure S11: **Inferring selection for lactase persistence**. Here we show the results of our sweep inference method (star-like approximation) for a number of test sites spaced 50Kb apart in the LCT region. The top panel shows the Composite Likelihood Ratio scores. Results in gray use data from both sides of the test site; results in black include only data from the intergenic region downstream of the focal rs4988235 mutation (centered at 0). The middle panel shows the estimated strength of selection (red) and age of the sweep (blue). Results using downstream data only are shown as solid lines whereas results using data from both sides are dashed. The bottom panel shows the genic (transcribed) content of the full genomic region we used in our analysis (http://genome.ucsc.edu).