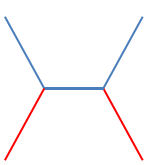
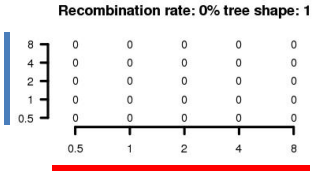
No Recombination

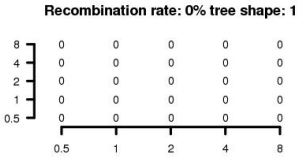
PhyMLE

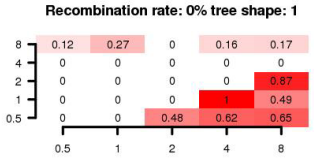
PhyMLA

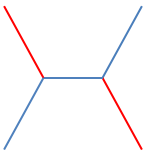
PhyMLS

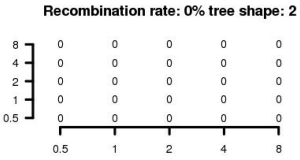


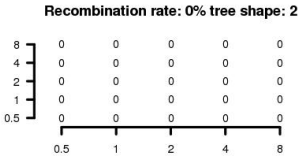


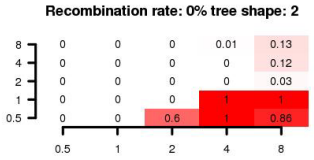


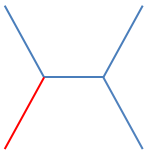


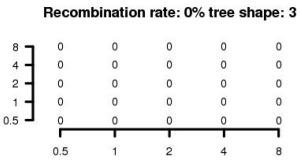


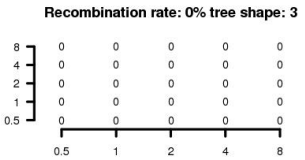


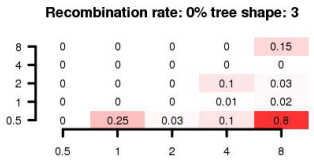


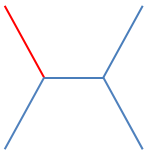


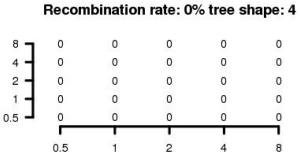


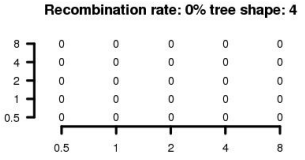


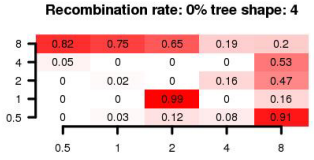


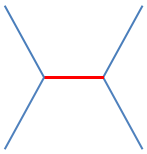


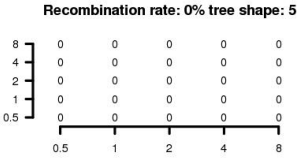


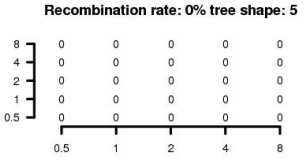


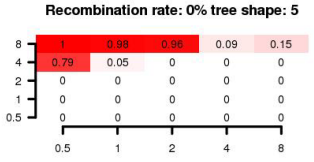


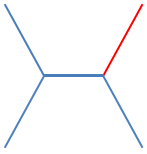


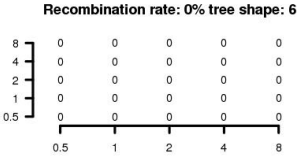


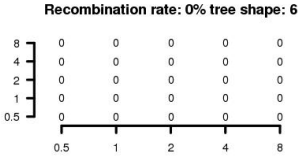


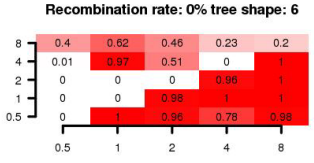


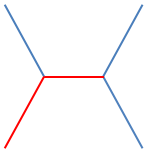


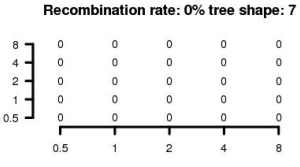


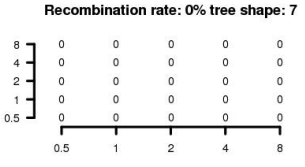


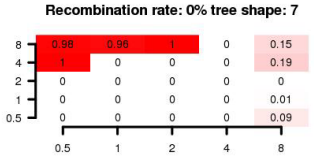


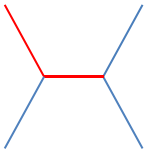


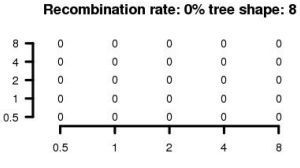


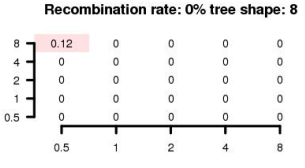


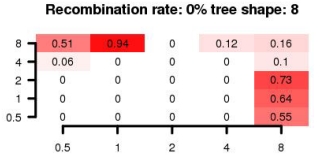


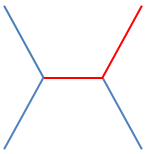


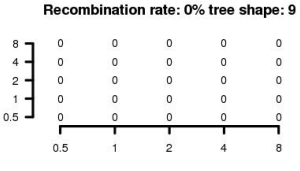


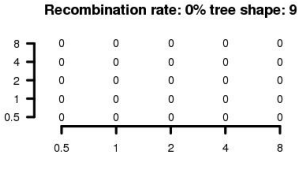


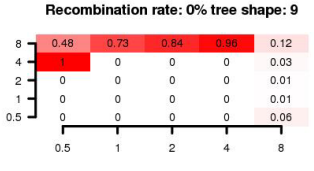


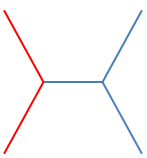


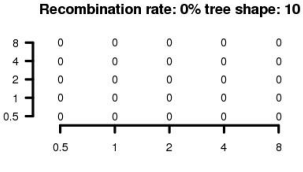


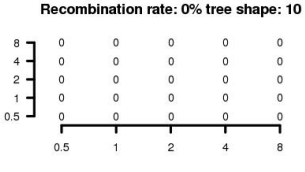


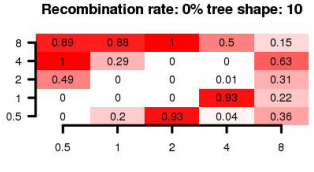






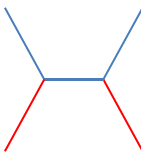


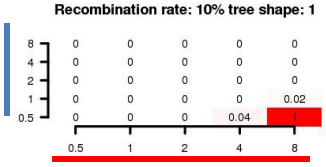


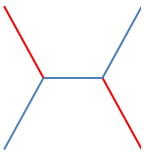


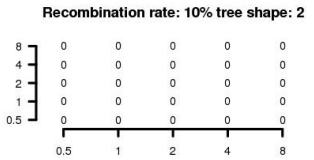
10% Recombination

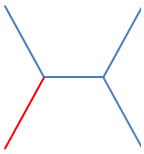
PhyMLE

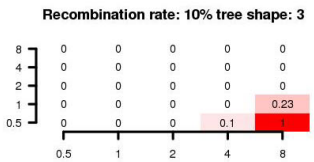


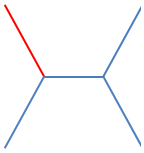


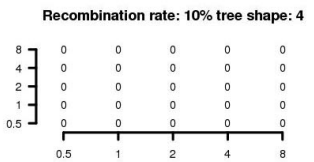


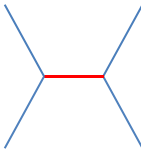


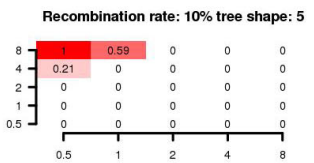


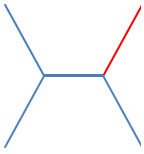


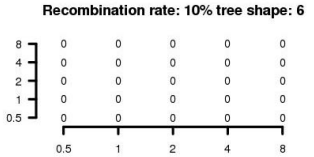


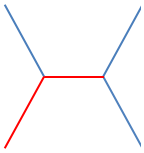


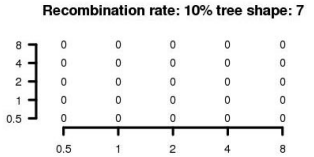


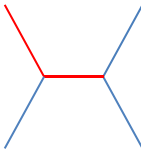


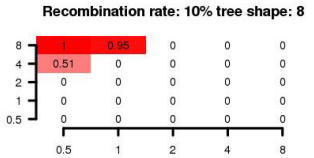


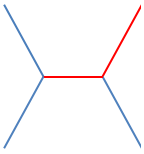


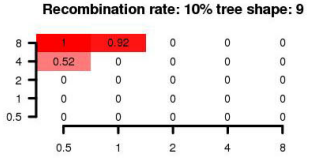


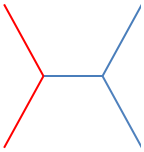


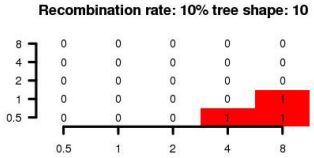


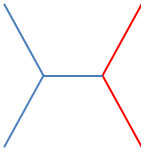


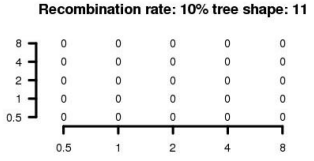




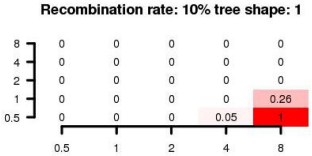


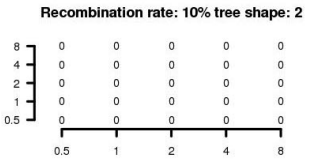


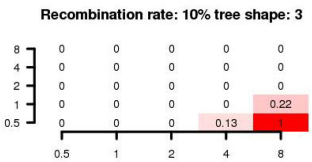


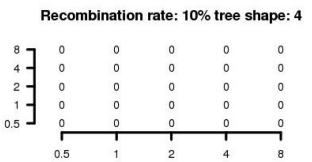


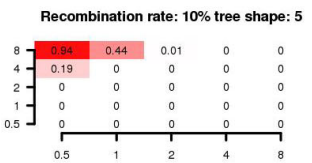
PhyMLA

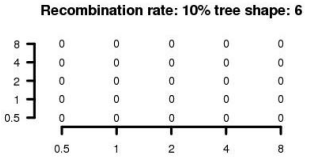


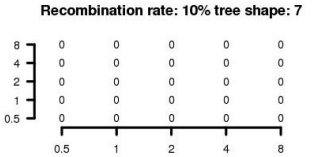


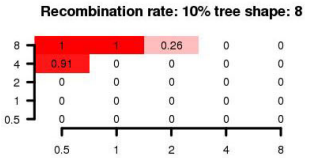


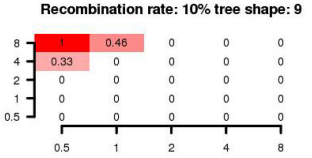


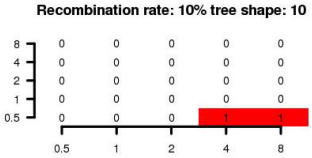


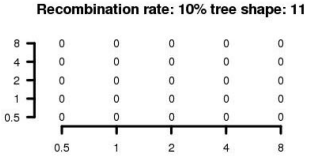




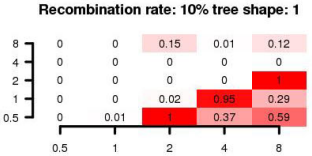


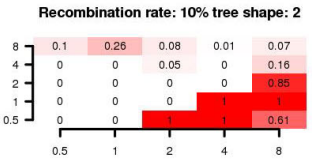


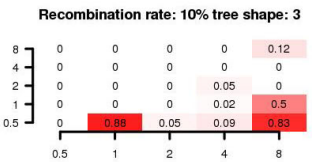


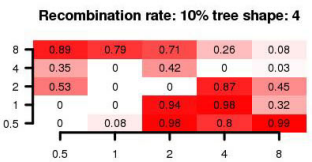


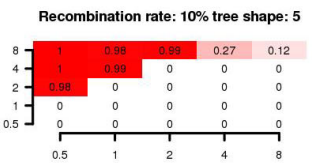
PhyMLS

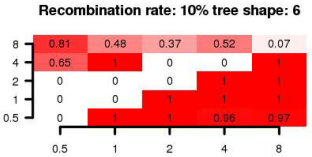


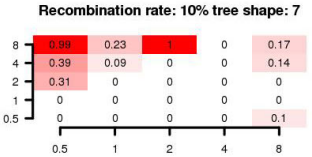




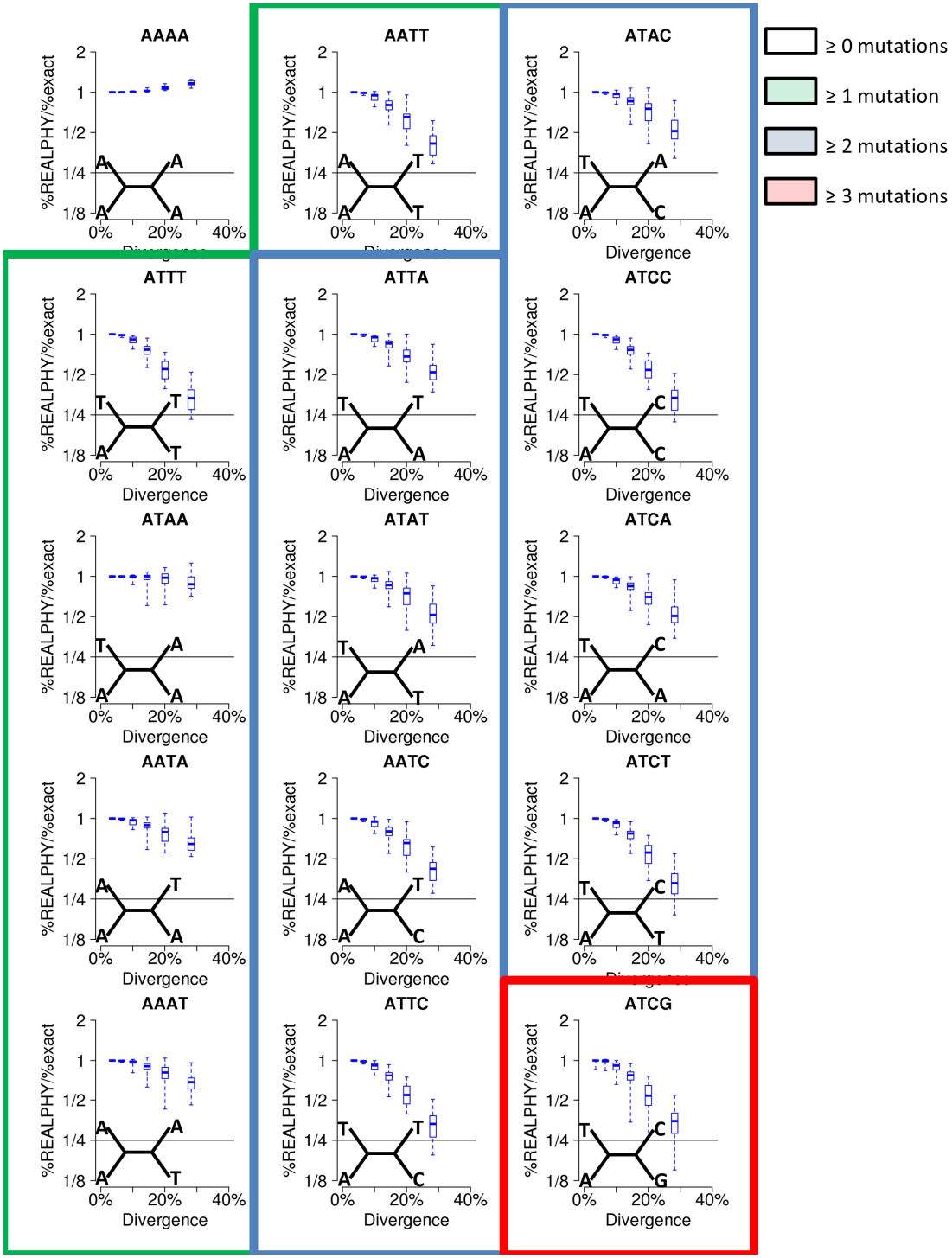






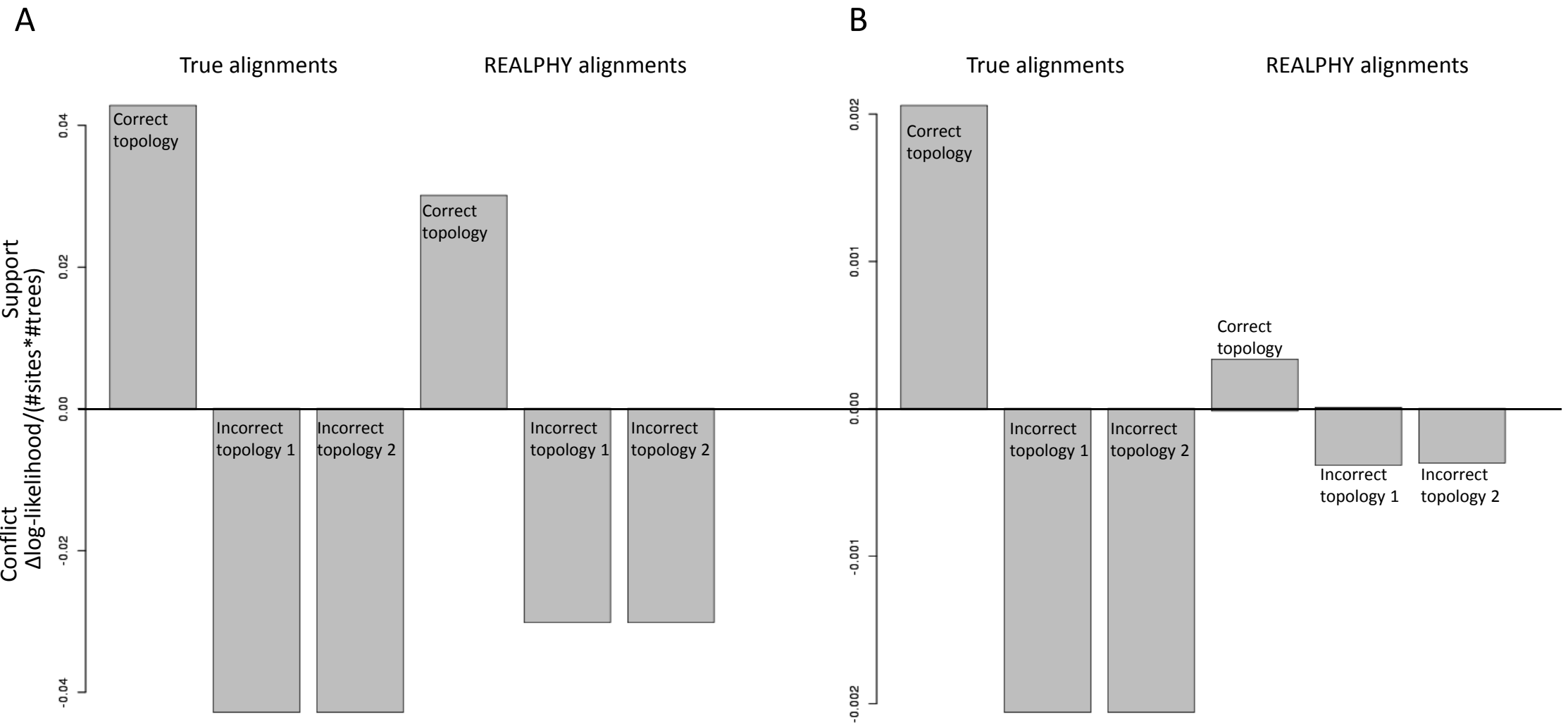


Supplementary Figure 1. Fraction of incorrectly reconstructed trees from a total of 100 replicates for all parameter combinations. Each column shows the result for a different type of alignment. The first column (PhyMLE) shows all trees reconstructed *via* PhyML (Guindon et al. 2010) from the exact alignment of the simulated sequences; the second (PhyMLA), PhyML trees based on all mapped sites; the third (PhyMLS), PhyML trees based only on extracted SNPs. The first page of the figure shows data without recombination between reference and cousin and the second page data that includes a recombination rate of 10%. Each row shows data for a different tree shape. Tree shapes are indicated on the left of the figure. Each individual heatmap shows the divergence of red branches across the x-axis and divergences across blue branches on the y-axis in percent.

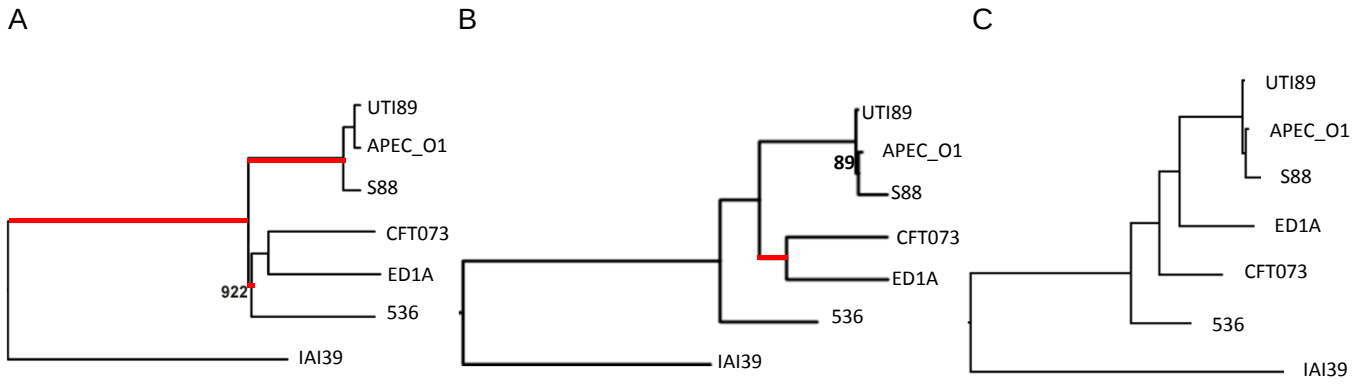


Supplementary Figure 2. Ratios of site class frequencies in REALPHY and true alignments, as a function of the total divergence within the phylogenetic tree. All datasets from all 550 parameter settings were stratified into six bins as a function of the total sequence divergence in the tree. Sequence divergence is defined as the proportion of variant alignment sites (not AAAA). For each of the 15 possible site classes, we then calculated site class ratios by dividing the proportion of the specific site class in REALPHY alignments by the proportion of this site

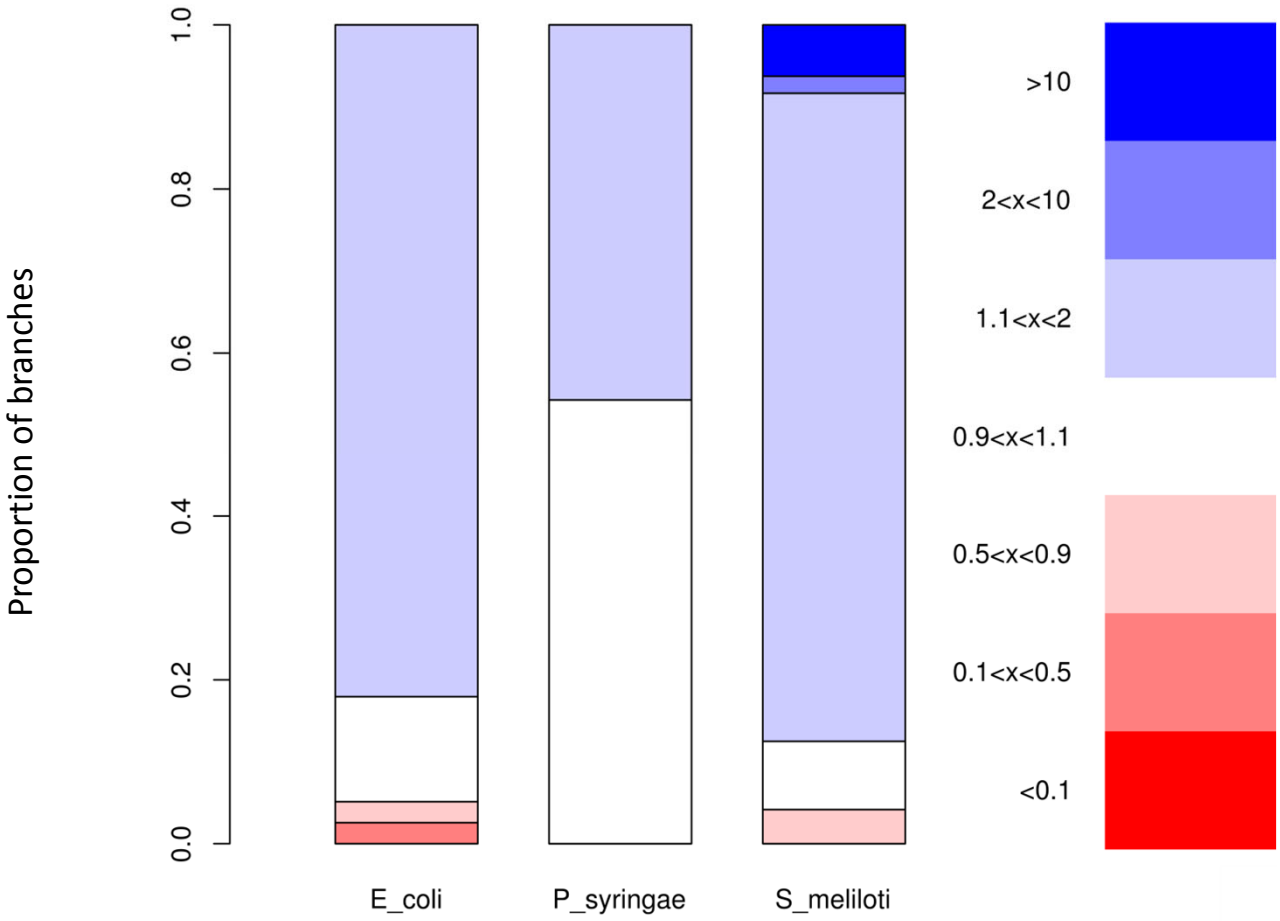
class in true alignments. Each panel corresponds to one site class (indicated at the top, and illustrated in the four taxon tree) and shows boxplots indicating quartiles, median, minimum and maximum site class ratios. All graphs show that the quality of REALPHY alignments decreases (values diverge from one) with increasing sequence divergence.



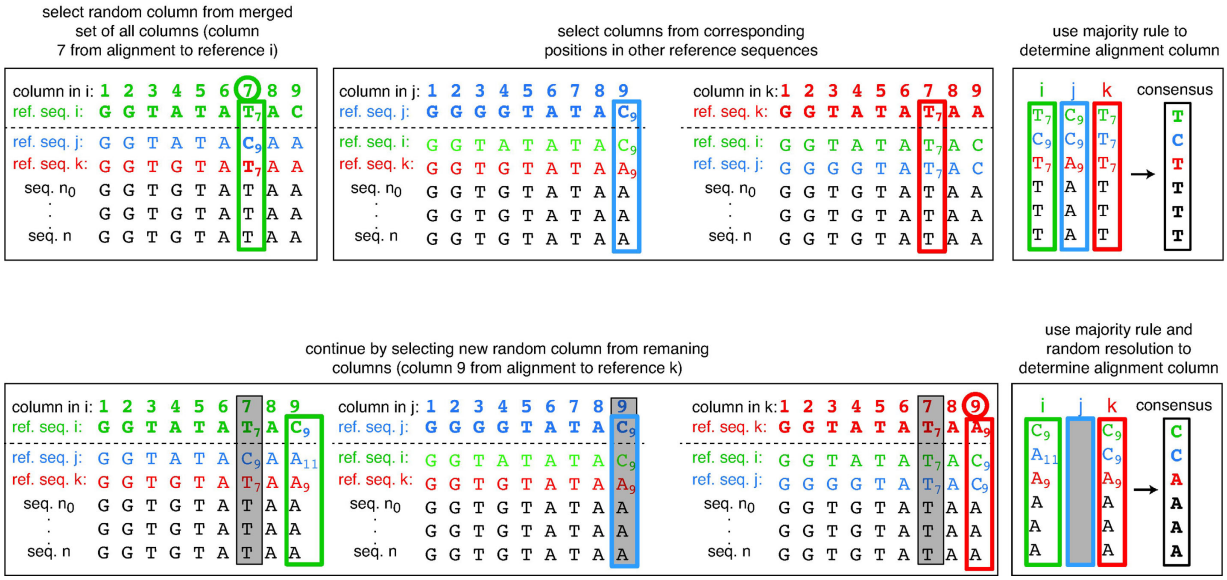
Supplementary Figure 3. PhyML log-likelihood differences per site and number of trees between correct and incorrect topologies for (*A*) all simulated four taxon trees without recombination and for (*B*) 100 repeats of tree shape 8 with 0.5% and 8% divergence. To measure the support and conflict of an alignment for a phylogeny we first determined the likelihood of the optimal phylogenetic tree for each of the three possible topologies for a four taxon tree. We then defined the support/conflict for a certain topology as the difference between the optimized log-likelihood of this topology and the maximum of the log-likelihoods optimized for the two remaining topologies. If this difference was positive then we added the value as support for the topology, if the difference was negative then we added it to the conflicting evidence. We also normalized the difference for sequence length. We applied this method to all simulated data sets without recombination (*A*) as well as for the problematic tree shape 8 with 8% and 0.5% divergence (*B*) in each case normalized for the number of trees in the subset (*i.e.* 27500 for (A) and 100 for (B)). As expected, the support for the mapped alignments is significantly decreased compared to the true alignments. This effect is most pronounced for the problematic parameter combination (*B*). Conflict bars for the correct topology and support bars for the incorrect topologies are too small to see.



Supplementary Figure 4. Phylogenetic trees of the *E. coli* B2 clade. (*A*) This phylogeny of the *E. coli* B2 clade is from Touchon *et al.* (2009). It is based on a 1,769,508 bp long alignment covering about 40% of the shortest genome. The phylogeny differs in three branch points from the phylogeny in *C* (red branches). (*B*) subtree of the *E. coli* phylogeny inferred from a merged REALPHY alignment of 20 *E. coli* strains and one *E. fergusonii* strain, with a total length of about 1,896,194 bp covering about 43% of the shortest genome. This phylogeny differs in one branch point from the phylogeny in *C* (red branch). (*C*) This phylogeny was inferred from a merged REALPHY alignment of all seven B2 strains with a total length of 3,793,038 bp or 76.8% of the shortest B2 genome. Due to the size of the alignment we assume that this phylogeny is likely to model the evolution of the B2 strains most accurately. The phylogenies in *B* and *C* were built using PhyML, with GTR as substitution model.



Supplementary Figure 5. The proportion of branches in three bacterial phylogenies inferred from SNP sites only whose relative branch lengths (branch length divided by the sum of all branch lengths in the phylogeny) differ to a certain degree (from more than 10 times shorter (dark red) to more than 10 times longer (dark blue)) from relative branch lengths inferred from complete REALPHY alignments (including non-polymorphic sites). The *E. coli* trees were inferred from a merged alignment of all 21 reference genomes. The *P. syringae* trees were inferred from a merged alignment of three reference genomes (*P. syringae pv. phaseolicola* 1448a*, P. syringae* B728a and *P. syringae pv. tomato* DC3000). The *S. meliloti* trees were inferred from a reference alignment to *S. meliloti* 1021 and *S. meliloti* Rm41. All trees were inferred *via* PhyML with the general time reversible (GTR) model of nucleotide evolution.



Supplementary Figure 6. Reference alignment merging. (1) Randomly select an alignment column from any of the reference alignments. (2) For each reference aligned within the selected column go to the position in the reference genome that the aligned site originated from and select the alignment column mapped to this position. (3) For all selected alignment columns calculate a consensus alignment column by applying a majority rule (most common nucleotide in each row gets selected for the new row in the consensus column) and add the consensus column to the merged alignment. (4) Remove the selected columns from all reference alignments, in order to ensure that no alignment column contributes multiple times to the merged alignment.