Color legend:

blue:

- Found on Illinois data bank at https://www.ideals.illinois.edu/collections/1451
- Find these files using the simple search option.

Green: found here on the GitHub.

Code

A pipeline is used for performing the binning step.

- The updated version of the code is available on GitHub: https://github.com/smirarab/binning
- The code that we used in the binning paper is available **binning-code.zip (https://www.ideals.illinois.edu/items/55536)**. Once you unzip the file, look at the README file for usage and installation guidelines.
- Note that this pipeline works on *nix-like systems (including MAC) but not on Windows; However, the main code to perform vertex coloring and to perform compatibility checks is in Java and can run on Windows. You just need to develop some gluing scripts if you are on Windows.

Simulated datasets

- The model species trees for 1X model condition: **avian-model-species.tre** and **mammalian-model-species.tre**. (for reduced or increased ILS model conditions, we simply multiply or divide the branch lengths by 2 or 5).
- Steps of the simulation are described in more detail in the **simulation** directory.
- In addition, we provide the definition of bins for all our supergene trees for our avian (**avian-bin-dfs.tar.bz2**) and mammalian (**mammals-bin-dfs.tar.bz2**) datasets. These files contain a `pairwise/R/[50/75]/bin..txt` file for each of the model conditions. These files are simple text files that give the gene ids put into each bin.
- We also provide the estimated species trees. You can download concatenation, unbinned MP-EST and binned MP-EST here for the avian (**avian-species-trees.tar**) and mammalian (**mammalian-species-trees.tar**) simulated datasets.
- The archive files noted in the table below contain
    1. simulated true gene trees
    2. simulated sequence data (alignments in fasta format)
    3. estimated gene trees and their bootstrap replicates

This table gives the names of the files corresponding to each condition.

Avian - 1X	**Avian - 0.5X**	**Avian - 2X**	**Mammalian - 1X**	**Mammalian - 0.5X**	**Mammalian - 2X**
avian-1X-truegt.tar.bz2 avian-1X-sequence.tar.bz2 avian-1X-estimated-genetrees.tar.bz2	avian-0.5X-truegt.tar.bz2 avian-0.5X-sequence.tar.bz2 avian-0.5X-estimated-genetrees.tar.bz2	avian-2X-truegt.tar.bz2 avian-2X-sequence.tar.bz2 avian-2X-estimated-genetrees.tar.bz2	mammalian-1X-truegt.tar.bz2 mammalian-1X-sequence.tar.bz2 mammalian-1X-estimated-genetrees.tar.bz2	mammalian-0.5X-truegt.tar.bz2 mammalian-0.5X-sequence.tar.bz2 mammalian-0.5X-estimated-genetrees.tar.bz2	mammalian-2X-truegt.tar.bz2 mammalian-2X-sequence.tar.bz2 mammalian-2X-estimated-genetrees.tar.bz2

Each model condition has a separate directory, with a name of the form:

```
[ILS]-[gene_count]-[alignment_length]
```

- ILS can be 1X, 0.5X, or 2X for both datasets.
- The gene_count can be 250, 500, 1000, or 2000, for avian, and 200, 400, or 800 for mammalian dataset.
- The alignment_length can be 250, 500, 1000, 1500, or true for avian and 500, 1000, or true for the mammalian dataset.

For avian 1X and mammalian 1X model conditions, we include sequence alignments only in 1X-1000-1500, and 1X-800-1000 directories respectively, and do not repeat them in the other ones. We can do this because various model conditions with 1X ILS use the same set of underlying simulated gene alignments and gene trees. For example, a subset of simulated genes used in the 1X-1000-500 model condition are used to create the 1X-500-500 condition. Thus, for model conditions that are missing alignments and trees in their directories, you can simply copy (or link) the alignments and trees from the corresponding gene (i.e. the same gene number) from the 1X-1000-1500 avian or the 1X-800-1000 mammalian directories. For model conditions with less than 1500 sites for avian or 1000 sites for mammalian, the alignments need to be trimmed to the first [alignment_length] sites. Thus to get alignments for 1X-1000-1000 model condition, the alignments from 1X-1000-15000 directory need to be trimmed to their first 1000 sites, discarding the final 500 sites.

Biological datasets

There are 5 biological datasets.

- The avian dataset is available at http://dx.doi.org/10.5524/101041.
- The mammalian dataset was provided to us by Song et al from their http://www.pnas.org/content/109/37/14942.long. We provide the alignments (from Song et al), the gene trees, and the supergene trees that we estimated on those alignments: **song-mammalian-bio.zip**
- The remaining three datasets are all from http://www.nature.com/nature/journal/v497/n7449/full/nature12130.html. The authors kindly provided to us both the alignments and their gene trees. We make the supergene alignments and trees (which we estimated) available in the files **salichos-bio.zip** and salichos-alignments.zip.