

Supplementary Material for:

**PhyloCoalSimulations: A simulator for network multispecies
coalescent models, including a new extension for the inheritance
of gene flow**

John Fogg, Elizabeth S. Allman, and Cécile Ané

Correlation from the Dirichlet process

Consider a given hybrid node with $k \geq 2$ parents, and let m be one of its k parental populations. We prove here that the correlation is $\rho = 1/(1 + \alpha)$ between Z_{i_1} and Z_{i_2} , where Z_i is equal to 1 if lineage i is inherited from parental population m and 0 otherwise. To simplify notations, let $\gamma = \gamma_m$ be the probability that a lineage is inherited from parent m . Then each Z_i has a Bernoulli distribution $Z_i \sim \mathcal{B}(\gamma)$ individually. Since the Dirichlet process is exchangeable, we simply need to consider the case $i_1 = 1$ and $i_2 = 2$. Then $\text{cor}(Z_1, Z_2) = (\mathbb{E}(Z_1 Z_2) - \gamma^2)/(\gamma(1 - \gamma))$ and the result follows easily from

$$\mathbb{E}(Z_1 Z_2) = \mathbb{P}(Z_1 = 1 \text{ and } Z_2 = 1) = \gamma \left(\frac{1}{1 + \alpha} + \frac{\alpha}{1 + \alpha} \gamma \right).$$

Validation of the distribution of quartet concordance factors

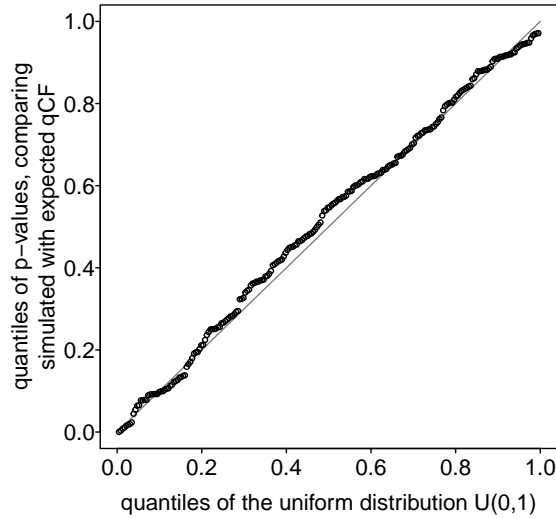


Figure S1: Distribution of 230 p-values from chi-square tests comparing qCFs observed in simulated gene trees (Fig. 2c, vertical axis) with qCFs expected from the network (Fig. 2c, horizontal axis). The distribution of p-values closely matches a uniform distribution, as expected.

Validation of the distribution of pairwise distances

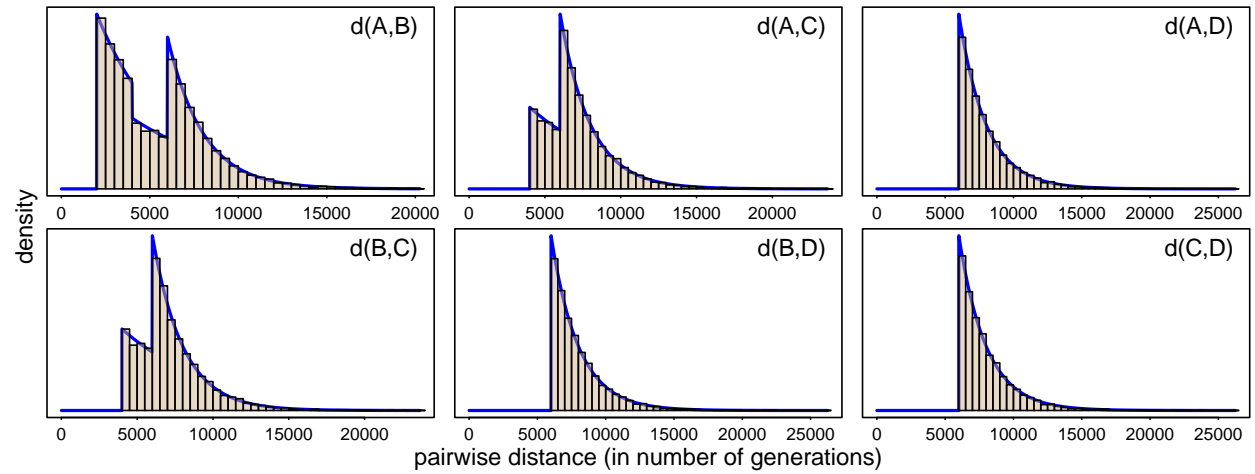


Figure S2: True distribution (blue curve) and distribution observed from 100,000 simulated gene trees (brown histogram) of the distance between pairs of taxa, under the coalescent with $\rho = 0$ on the asymmetric tree from Figure 3 of Allman et al. (2023): $((A:1000, B:1000):1000, C:2000):1000, D:3000$, and population size of 2000, 3000 and 1000 in lineages ancestral to AB , ABC and the root, respectively.

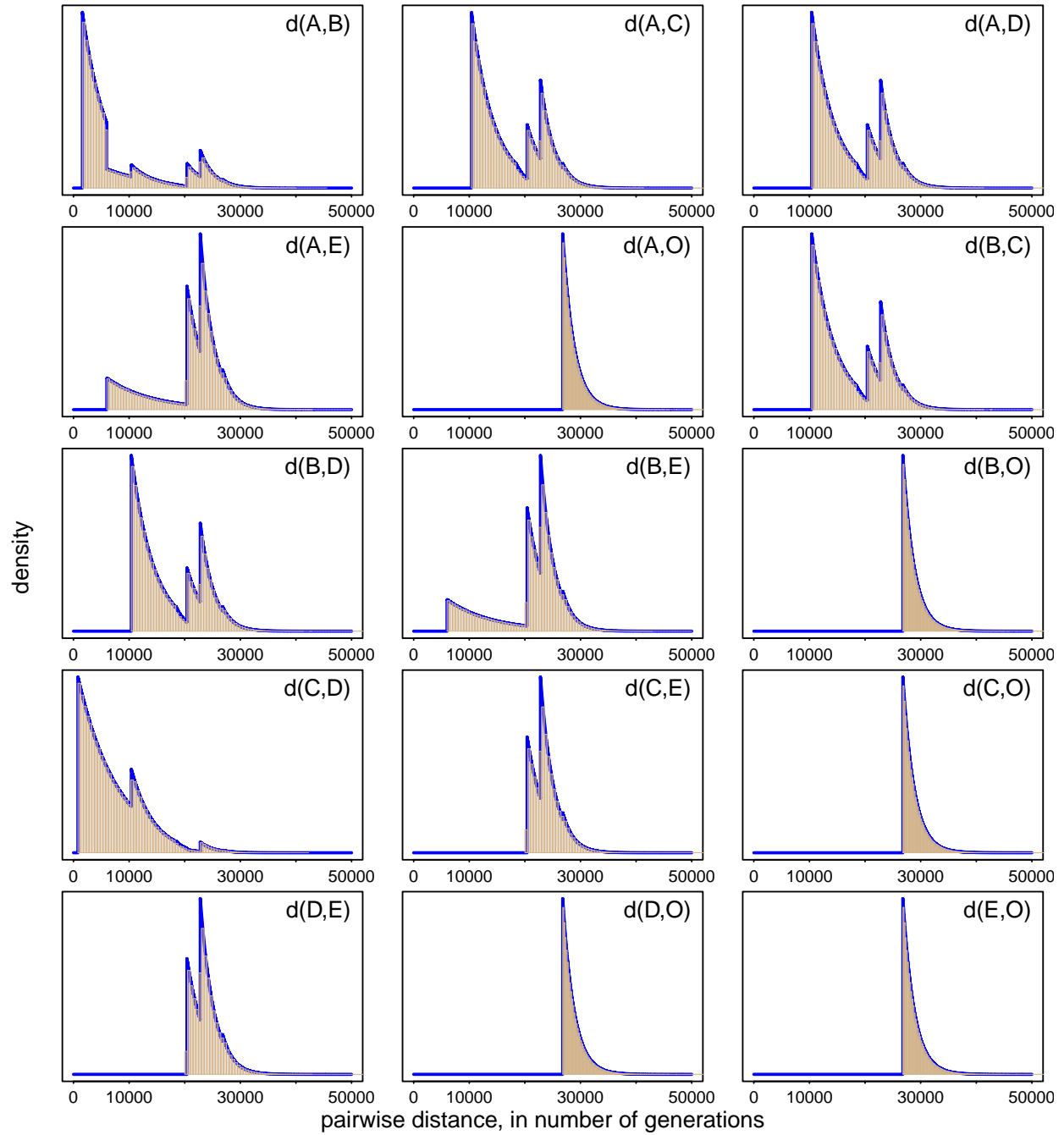


Figure S3: True distribution (blue curve) and distribution observed from 100,000 simulated gene trees (brown histogram) of the distance between all 15 pairs of taxa, under the coalescent with $\rho = 0$ on the level-2 network in Figure 3. For ease of comparison, six of the histograms in Figure 4 are repeated here.

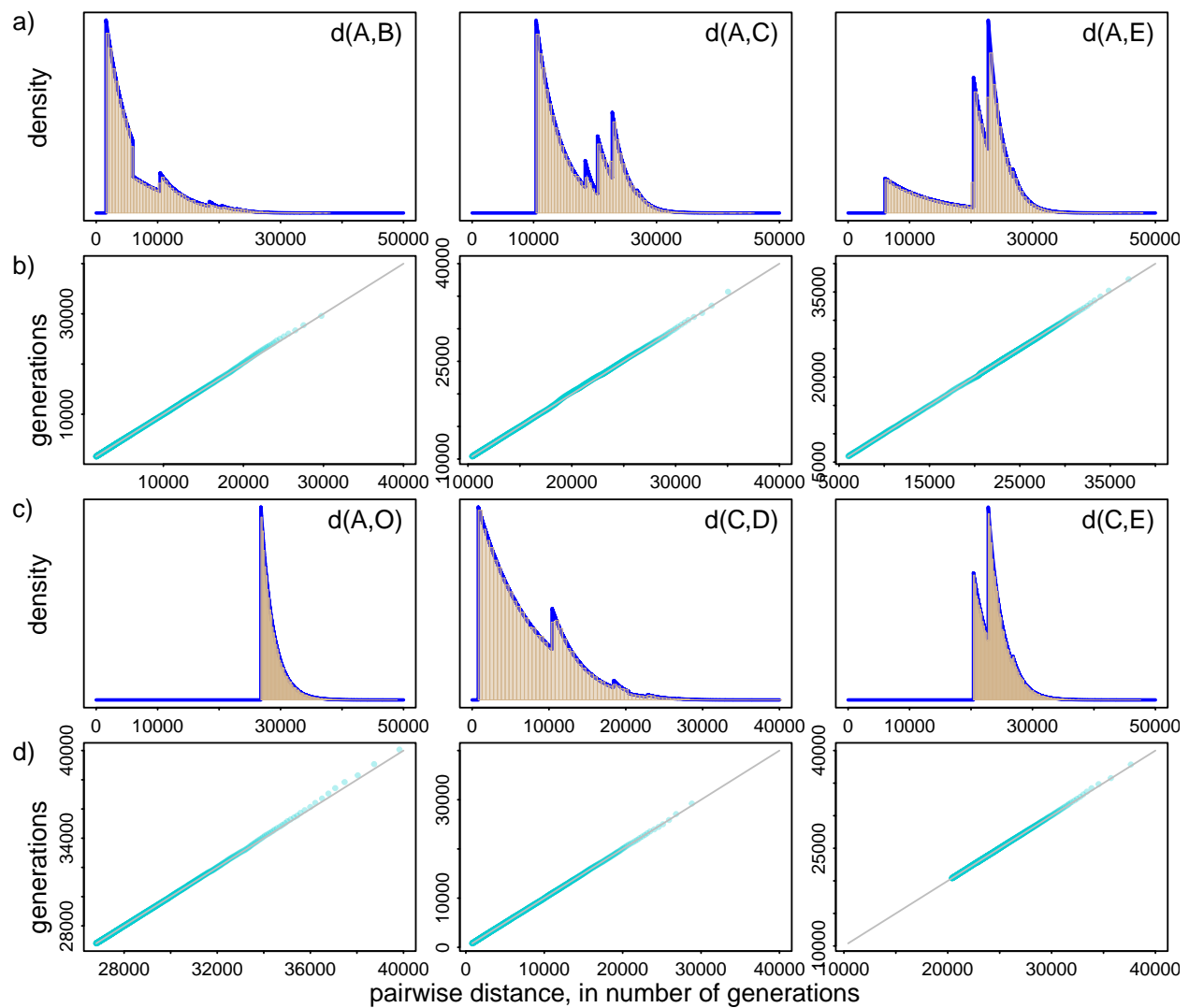


Figure S4: Comparison of the simulated and true distributions of pairwise distances under the coalescent on the level-2 network in Figure 3, for 6 pairs as in Figure 4 but with common inheritance: $\rho = 1$.

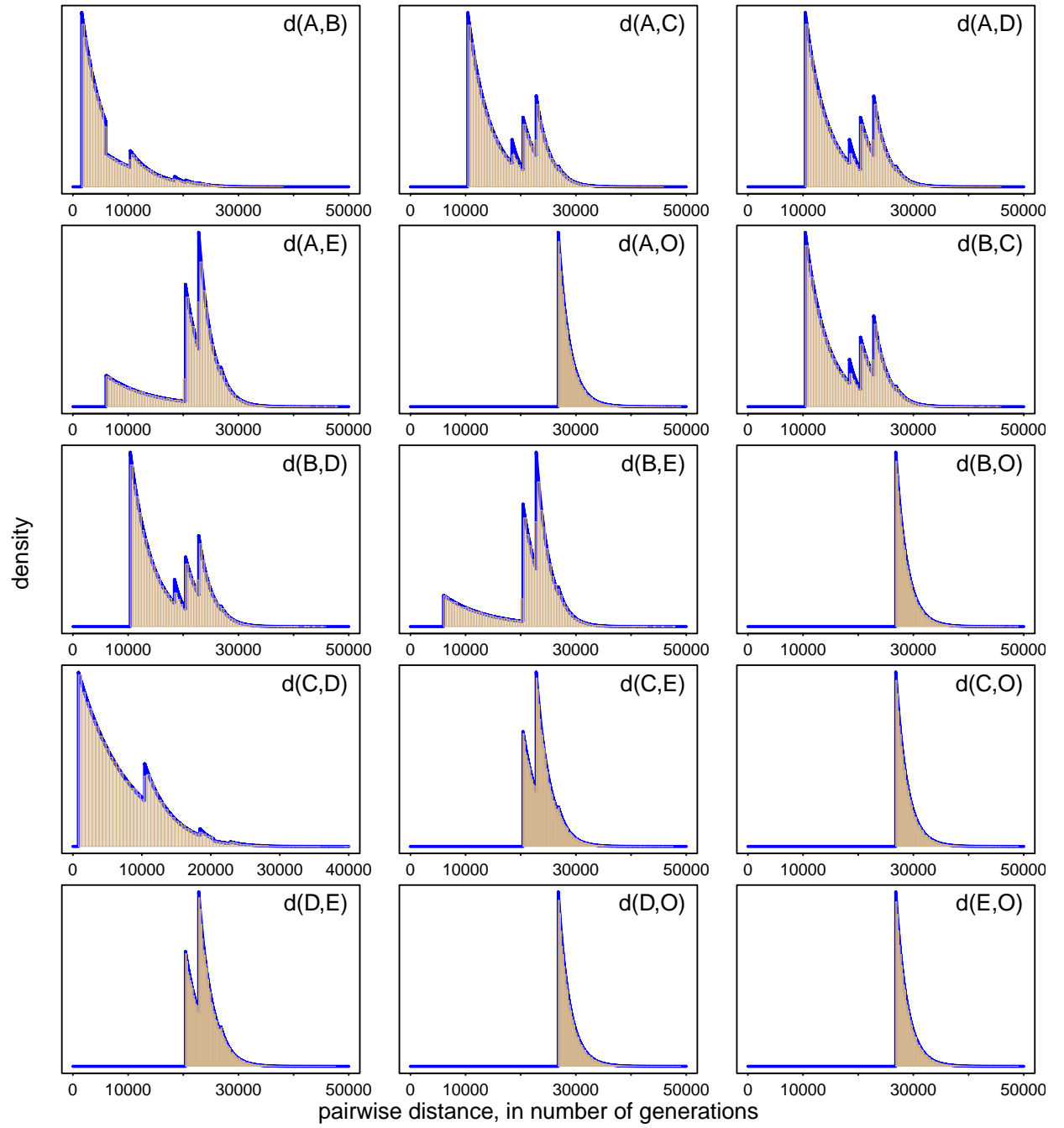


Figure S5: Comparison of the simulated and true distributions of pairwise distances under the coalescent on the level-2 network in Figure 3, for all 15 pairs of taxa as in Figure S3 but with common inheritance: $\rho = 1$.

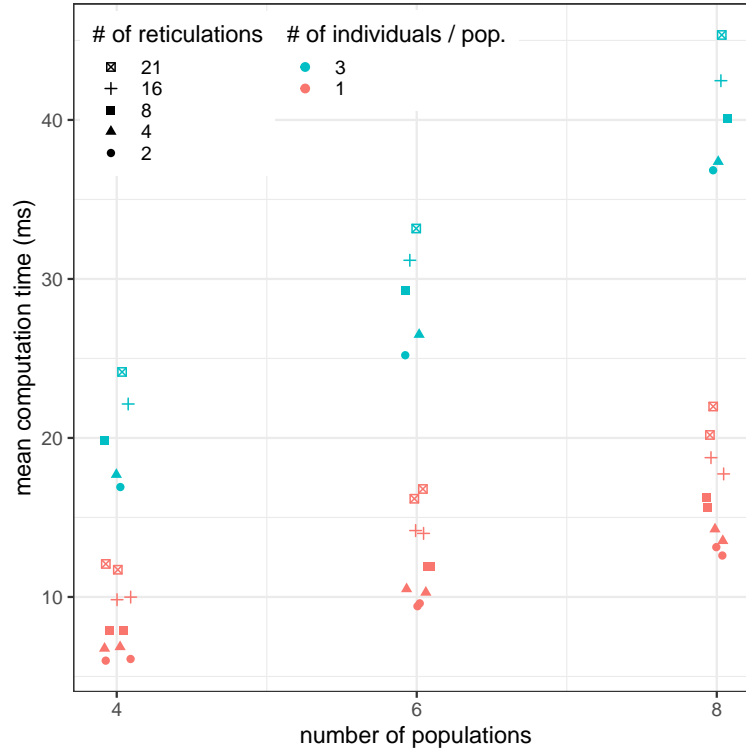


Figure S6: Average computing time to simulate 1,000 gene trees with `PhyloCoalsSimulations` using a single processor on a 3.5GHz Linux machine. For each number of reticulations (distinguished by point shapes), one 8-taxon network was used. From this network, taxa were pruned to obtain a 6-taxon and a 4-taxon network with the same number of reticulations. Simulated gene trees had 1 individual per population (red) or 3 individuals per population (teal). For 1 individual per population, benchmarks were repeated after dividing all edge lengths by 3, to increase the level of incomplete lineage sorting and the gene tree embedding complexity. The 2 settings of branch lengths are shown with separate markers, for each number of reticulations and number of populations (with 1 individual per population).

References

E. S. Allman, H. D. Baños, and J. A. Rhodes. Testing multispecies coalescent simulators using summary statistics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(02):1613–1618, mar 2023. ISSN 1557-9964. doi: 10.1109/TCBB.2022.3177956.