

Reference genome choice and filtering thresholds jointly influence phylogenomic analyses

JESSICA A. RICK, CHAD D. BROCK, ALEXANDER L. LEWANSKI, JIMENA GOLCHER-BENAVIDES, AND CATHERINE E. WAGNER

**Corresponding author: JA Rick, University of Arizona, 1064 E Lowell St, Tucson, AZ 85719 USA; jrick@arizona.edu*

SUPPLEMENTARY MATERIAL

Supplemental methods

Reference genome for simulations.— To create a reference genome to use as a base for our TreeToReads simulations, we started with the *Lates calcarifer* reference genome (Vij et al., 2016), as this was the most complete of the four reference genomes used in our empirical analyses. We randomly sampled 5,000bp windows without replacement using a custom bash script. For each random segment, we replaced any ambiguous bases (N) with a random base (A, C, G, or T), with equal weighting given to each option.

Genotyping-by-sequencing for Lake Tanganyika tropheine cichlids.— We sampled individuals in the cichlid tribe Tropheini in the Kigoma region of Tanzania in 2002, 2005, 2007, 2012, 2015, 2016, 2017, and 2018. Briefly, we collected individuals using gill nets while snorkeling in the rocky littoral zone at a 1–10 m depth. We took a fin clip from each individual for genetic analyses, which was stored in either DMSO-EDTA or 95% ethanol prior to DNA extraction. We retained all individuals as vouchers, which are deposited at the Cornell University Museum of Vertebrates (2002, 2005, and 2007 individuals) and the University of Wyoming Museum of Vertebrates (2015, 2016, 2017, and 2018 individuals).

We extracted DNA from the fin clips using DNeasy Blood & Tissue kits (Qiagen, Inc.) and prepared genomic libraries for high-throughput DNA sequencing following the GBS protocol described in Parchman et al. (2012). Briefly, we fragmented DNA using EcoRI and MseI restriction enzymes and then ligated a unique 8 – 10 base pair (bp) barcode to each individual’s DNA. We used polymerase chain reaction (PCR) to amplify the restriction/ligation products (two independent replicates per individual) and then combined the PCR products to create the final libraries. Prior to sequencing, the libraries were size-selected for 250 – 400 bp fragments using BluePippin (Sage Science). Sequencing was completed on Illumina HiSeq 2500 and 4000 platforms (100 bp single-end) at the University of Texas Genome Sequencing and Analysis Facility (Austin, Texas, USA) and the University of Oregon Genomics and Cell Characterization Core Facility (Eugene, Oregon, USA). The individuals in this project were included in libraries containing samples of other cichlid species as part of a larger sequencing effort. Each library contained ~100 individuals and was sequenced in its own lane.

With the raw sequence data, we first assigned reads to individuals and subsequently removed the barcode sequences using a custom perl script. We then aligned the reads to the *Pundamilia nyererei* and *Oreochromis niloticus* reference genomes (Brawand et al., 2015) using bwa v0.7.17 (v0.7.17 Li and Durbin, 2009) with default settings to produce bam

alignment files, which were used for variant calling in downstream analyses.

REFERENCES

- Brawand, D., C. E. Wagner, Y. I. Li, M. Malinsky, I. Keller, S. Fan, O. Simakov, A. Y. Ng, Z. W. Lim, E. Bezault, J. Turner-Maier, J. Johnson, R. Alcazar, H. J. Noh, P. Russell, B. Aken, J. Alföldi, C. Amemiya, N. Azzouzi, J. F. Baroiller, F. Barloy-Hubler, A. Berlin, R. Bloomquist, K. L. Carleton, M. A. Conte, H. D’Cotta, O. Eshel, L. Gaffney, F. Galibert, H. F. Gante, S. Gnerre, L. Greuter, R. Guyon, N. S. Haddad, W. Haerty, R. M. Harris, H. A. Hofmann, T. Hourlier, G. Hulata, D. B. Jaffe, M. Lara, A. P. Lee, I. MacCallum, S. Mwaiko, M. Nikaido, H. Nishihara, C. Ozouf-Costaz, D. J. Penman, D. Przybylski, M. Rakotomanga, S. C. Renn, F. J. Ribeiro, M. Ron, W. Salzburger, L. Sanchez-Pulido, M. E. Santos, S. Searle, T. Sharpe, R. Swofford, F. J. Tan, L. Williams, S. Young, S. Yin, N. Okada, T. D. Kocher, E. A. Miska, E. S. Lander, B. Venkatesh, R. D. Fernald, A. Meyer, C. P. Ponting, J. T. Streelman, K. Lindblad-Toh, O. Seehausen, and F. Di Palma. 2015. The genomic substrate for adaptive radiation in African cichlid fish. *Nature* 513:375–381.
- Li, H. and R. Durbin. 2009. Fast and accurate short-read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
- Parchman, T. L., Z. Gompert, J. Mudge, F. D. Schilkey, C. W. Benkman, and C. A. Buerkle. 2012. Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology* 21:2991–3005.
- Vij, S., H. Kuhl, I. S. Kuznetsova, A. Komissarov, A. A. Yurchenko, P. Van Heusden, S. Singh, N. M. Thevasagayam, S. R. S. Prakki, K. Purushothaman, J. M. Saju, J. Jiang, S. K. Mbandi, M. Jonas, A. Hin Yan Tong, S. Mwangi, D. Lau, S. Y. Ngoh, W. C. Liew, X. Shen, L. S. Hon, J. P. Drake, M. Boitano, R. Hall, C. S. Chin, R. Lachumanan, J. Korlach, V. Trifonov, M. Kabilov, A. Tupikin, D. Green, S. Moxon, T. Garvin, F. J. Sedlazeck, G. W. Vulture, G. Gopalapillai, V. Kumar Katneni, T. H. Noble, V. Scaria, S. Sivasubbu, D. R. Jerry, S. J. O’Brien, M. C. Schatz, T. Dalmay, S. W. Turner, S. Lok, A. Christoffels, and L. Orbán. 2016. Chromosomal-Level Assembly of the Asian Seabass Genome Using Long Sequence Reads and Multi-layered Scaffolding. *PLoS Genetics* 12:1–35.

Supplemental tables and figures

Table S1. Model coefficients from linear mixed models, including parameter estimates (Est), minimum 95% confidence interval (Min CI), upper 95% confidence interval (Max CI), and significance at $\alpha = 0.05$ (* indicates $p < 0.05$). Parameters are listed for models for all RAxML trees, as well as the three ILS levels individually. Predictor variables in each model are average distance from the ingroup taxa to the reference genome (avg_dxy), minor allele count threshold (maf), missing data threshold (missing), and the interaction among these terms, with simulation number as a random effect.

Variable	All Trees				Low ILS				Medium ILS				High ILS			
	Est	Min CI	Max CI	Sig	Est	Min CI	Max CI	Sig	Est	Min CI	Max CI	Sig	Est	Min CI	Max CI	Sig
Standardized ingroup gamma																
avg_dxy	1.463	0.217	2.711	*	0.268	-0.647	1.182		0.203	-1.169	1.578		0.030	-1.238	1.300	
maf	0.815	0.760	0.870	*	0.611	0.566	0.656	*	0.937	0.880	0.995	*	0.851	0.799	0.903	*
missing	-0.371	-0.888	0.146		-0.282	-0.704	0.140		-0.451	-0.987	0.086		-0.378	-0.867	0.111	
avg_dxy:maf	-0.089	-0.280	0.102		-0.033	-0.173	0.108		-0.013	-0.223	0.197		-0.013	-0.208	0.181	
avg_dxy:missing	-0.005	-1.800	1.791		-0.170	-1.486	1.146		0.006	-1.969	1.980		0.172	-1.653	1.996	
Robinson-Foulds distance to true tree																
avg_dxy	-0.015	-0.035	0.005		0.005	-0.017	0.026		0.003	-0.024	0.029		-0.025	-0.061	0.012	
maf	0.013	0.012	0.014	*	0.016	0.015	0.017	*	0.012	0.011	0.013	*	0.012	0.010	0.013	*
missing	0.002	-0.007	0.010		0.001	-0.009	0.010		0.002	-0.009	0.012		0.001	-0.013	0.015	
avg_dxy:maf	0.003	0.000	0.006	*	0.000	-0.003	0.003		0.002	-0.002	0.006		0.007	0.002	0.013	*
avg_dxy:missing	-0.006	-0.035	0.023		-0.012	-0.043	0.018		-0.012	-0.050	0.027		0.014	-0.039	0.066	
Standardized ingroup Colless imbalance																
avg_dxy	0.019	0.004	0.034	*	0.006	-0.008	0.021		0.010	-0.019	0.038		0.032	0.005	0.059	*
maf	0.010	0.009	0.010	*	0.007	0.006	0.008	*	0.009	0.008	0.011	*	0.012	0.011	0.013	*
missing	-0.004	-0.011	0.002		-0.001	-0.008	0.006		-0.005	-0.016	0.006		-0.007	-0.018	0.003	
avg_dxy:maf	-0.002	-0.005	0.000	*	0.000	-0.002	0.002		-0.001	-0.005	0.004		-0.005	-0.009	-0.001	*
avg_dxy:missing	-0.001	-0.023	0.020		-0.006	-0.028	0.015		0.002	-0.039	0.043		0.000	-0.039	0.039	

Table S2. Model coefficients from linear mixed models for subsampled data sets, including parameter estimates (Est), minimum 95% confidence interval (Min CI), upper 95% confidence interval (Max CI), and significance at $\alpha = 0.05$ (* indicates $p < 0.05$). Parameters are listed for models for all trees, as well as the three ILS levels individually. Predictor variables in each model are average distance from the ingroup taxa to the reference genome (avg_dxy), minor allele count threshold (maf), missing data threshold (missing), and the interaction among these terms, with simulation number as a random effect.

Variable	All Trees				Low ILS				Medium ILS				High ILS			
	Est	Min CI	Max CI	Sig	Est	Min CI	Max CI	Sig	Est	Min CI	Max CI	Sig	Est	Min CI	Max CI	Sig
Standardized ingroup gamma																
avg_dxy	0.729	0.236	1.222	*	0.029	-0.435	0.493		0.313	-0.425	1.051		-0.074	-0.699	0.552	
maf	0.755	0.739	0.771	*	0.254	0.238	0.271	*	0.931	0.908	0.953	*	0.862	0.843	0.881	*
missing	-0.449	-0.645	-0.253	*	-0.034	-0.238	0.171		-0.590	-0.872	-0.309	*	-0.544	-0.780	-0.308	*
avg_dxy:maf	-0.127	-0.183	-0.071	*	0.021	-0.032	0.073		-0.075	-0.158	0.009		0.024	-0.047	0.095	
avg_dxy:missing	0.115	-0.579	0.809		-0.020	-0.673	0.632		0.036	-1.001	1.074		0.044	-0.836	0.924	
Robinson-Foulds distance to true tree																
avg_dxy	-0.014	-0.030	0.001		-0.012	-0.030	0.007		0.016	-0.006	0.039		-0.005	-0.031	0.020	
maf	0.006	0.006	0.007	*	0.015	0.014	0.016	*	0.005	0.005	0.006	*	0.002	0.002	0.003	*
missing	-0.005	-0.011	0.002		-0.003	-0.012	0.005		-0.001	-0.010	0.007		-0.009	-0.018	0.001	
avg_dxy:maf	0.003	0.001	0.005	*	0.001	-0.001	0.003		0.000	-0.002	0.003		0.002	-0.001	0.005	
avg_dxy:missing	0.003	-0.019	0.025		0.005	-0.022	0.031		-0.011	-0.043	0.020		0.016	-0.021	0.052	
Standardized ingroup Colless imbalance																
avg_dxy	-0.001	-0.011	0.009		0.007	-0.005	0.020		-0.010	-0.028	0.008		-0.001	-0.019	0.016	
maf	0.009	0.009	0.010	*	0.009	0.009	0.009	*	0.010	0.010	0.011	*	0.009	0.008	0.009	*
missing	-0.005	-0.009	-0.001	*	0.002	-0.004	0.007		-0.007	-0.014	0.000		-0.008	-0.015	-0.002	*
avg_dxy:maf	0.002	0.001	0.003	*	0.000	-0.001	0.002		0.004	0.002	0.006	*	0.001	-0.001	0.003	
avg_dxy:missing	-0.005	-0.019	0.010		-0.012	-0.030	0.006		-0.008	-0.034	0.017		0.002	-0.023	0.027	

Table S3. Model coefficients from linear mixed models for the empirical Tropheine and *Lates* data sets, including parameter estimates (Est), minimum 95% confidence interval (Min CI), upper 95% confidence interval (Max CI), and significance at $\alpha = 0.05$. Parameters are listed for models for all trees, as well as the three ILS levels individually. Predictor variables in each model are reference genome choice (ref), minor allele count threshold (mac), missing data threshold (missing), and the interaction among these terms, with iteration number as a random effect. Reference genome choice (ref) was coded as 0=outgroup reference and 1=ingroup reference.

Variable	Tropheine Topologies				Lates Topologies			
	Est	Min CI	Max CI	Sig	Est	Min CI	Max CI	Sig
Ingroup gamma								
ref	0.111	-0.037	0.259	*	0.030	-0.390	0.450	
mac	0.262	0.246	0.278	*	-0.254	-0.299	-0.208	*
missing	-1.204	-1.357	-1.051	*	-0.345	-0.773	0.082	
ref:mac	-0.025	-0.048	-0.002	*	0.021	-0.043	0.086	
ref:missing	0.420	0.207	0.634	*	0.161	-0.444	0.766	
Ingroup tree height								
ref	0.013	-0.007	0.034		-0.006	-0.034	0.022	
mac	0.022	0.019	0.024	*	-0.032	-0.035	-0.029	*
missing	-0.018	-0.039	0.004		-0.062	-0.091	-0.033	*
ref:mac	-0.020	-0.024	-0.017	*	0.000	-0.005	0.004	
ref:missing	0.032	0.002	0.062	*	-0.008	-0.049	0.032	
Ingroup Colless imbalance								
ref	0.020	-0.025	0.064		-0.011	-0.031	0.009	
mac	-0.011	-0.016	-0.006	*	-0.001	-0.003	0.001	
missing	-0.224	-0.270	-0.178	*	-0.009	-0.029	0.012	
ref:mac	0.002	-0.005	0.009		0.002	-0.001	0.005	
ref:missing	0.006	-0.059	0.070		0.002	-0.027	0.031	

Table S4. Model coefficients from linear mixed models for **ASTRAL** trees, including parameter estimates (Est), minimum 95% confidence interval (Min CI), upper 95% confidence interval (Max CI), and significance at $\alpha = 0.05$ (* indicates $p < 0.05$). Parameters are listed for models for all trees, as well as the three ILS levels individually. Predictor variables in each model are average distance from the ingroup taxa to the reference genome (avg.dxy), minor allele count threshold (mac), missing data threshold (missing), and the interaction among these terms, with simulation number as a random effect.

Variable	All Trees				Low ILS				Medium ILS				High ILS			
	Est	Min CI	Max CI	Sig	Est	Min CI	Max CI	Sig	Est	Min CI	Max CI	Sig	Est	Min CI	Max CI	Sig
Robinson-Foulds distance to true tree																
avg.dxy	-0.007	-0.022	0.009		-0.011	-0.036	0.015		-0.001	-0.022	0.020		0.001	-0.018	0.019	
maf	0.000	-0.001	0.000		0.000	-0.001	0.002		0.000	-0.001	0.001		-0.001	-0.002	0.000	*
missing	-0.008	-0.014	-0.002	*	-0.015	-0.027	-0.003	*	-0.006	-0.015	0.002		-0.002	-0.009	0.005	
avg.dxy:maf	0.001	-0.002	0.003		0.000	-0.004	0.003		-0.001	-0.004	0.002		0.004	0.001	0.007	*
avg.dxy:missing	-0.002	-0.024	0.020		0.018	-0.019	0.054		0.000	-0.031	0.030		-0.031	-0.058	-0.005	*
Standardized ingroup Colless imbalance																
avg.dxy	-0.018	-0.042	0.006		-0.021	-0.054	0.012		0.003	-0.028	0.035		-0.014	-0.046	0.019	
maf	0.019	0.018	0.020	*	0.028	0.027	0.030	*	0.016	0.015	0.018	*	0.014	0.013	0.016	*
missing	0.013	0.003	0.023	*	0.016	0.001	0.031	*	0.006	-0.006	0.019		0.019	0.007	0.032	*
avg.dxy:maf	0.007	0.003	0.011	*	0.002	-0.003	0.007		0.002	-0.003	0.007		0.009	0.005	0.014	*
avg.dxy:missing	0.000	-0.035	0.035		0.010	-0.037	0.058		-0.004	-0.049	0.041		-0.017	-0.064	0.030	

Table S5. Model coefficients from linear mixed models for **ASTRAL** trees for the empirical datasets, including parameter estimates (Est), minimum 95% confidence interval (Min CI), upper 95% confidence interval (Max CI), and significance at $\alpha = 0.05$ (* indicates $p < 0.05$). Parameters are listed for models for the Tropheine trees and *Lates* trees. Predictor variables in each model are reference genome choice (ref), minor allele count threshold (mac), missing data threshold (missing), and the interaction among these terms, with iteration number as a random effect. Reference genome choice (ref) was coded as 0=outgroup reference and 1=ingroup reference.

Variable	Tropheine Topologies				Lates Topologies			
	Est	Min CI	Max CI	Sig	Est	Min CI	Max CI	Sig
Ingroup Colless imbalance								
intINT	-0.025	-0.054	0.004		-0.051	-0.077	-0.026	*
maf	0.001	-0.002	0.005		0.015	0.012	0.017	*
missing	-0.127	-0.157	-0.098	*	-0.048	-0.074	-0.022	*
intINT:maf	-0.002	-0.006	0.003		0.004	0.000	0.007	
intINT:missing	-0.007	-0.049	0.035		-0.047	-0.084	-0.009	*

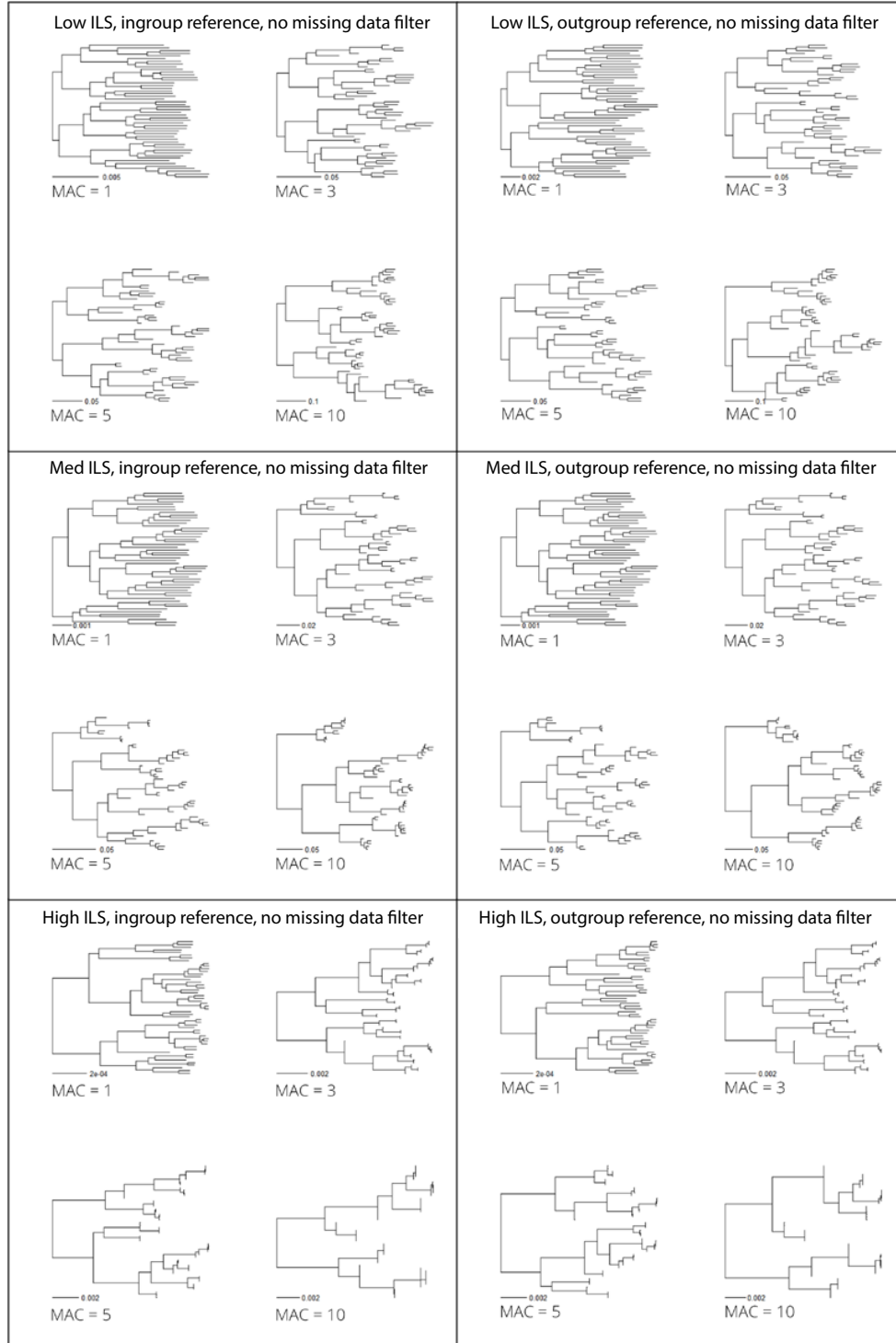


Fig. S1. Examples of inferred trees for a single simulation for high, medium, and low ILS species trees, where data were aligned to an ingroup (left) versus outgroup (right) reference genome. Trees are shown for no missing data cutoff and a minor allele count cutoff of 1, 3, 5, and 10, to demonstrate the changes in the trees that are summarized through the γ and imbalance statistics.

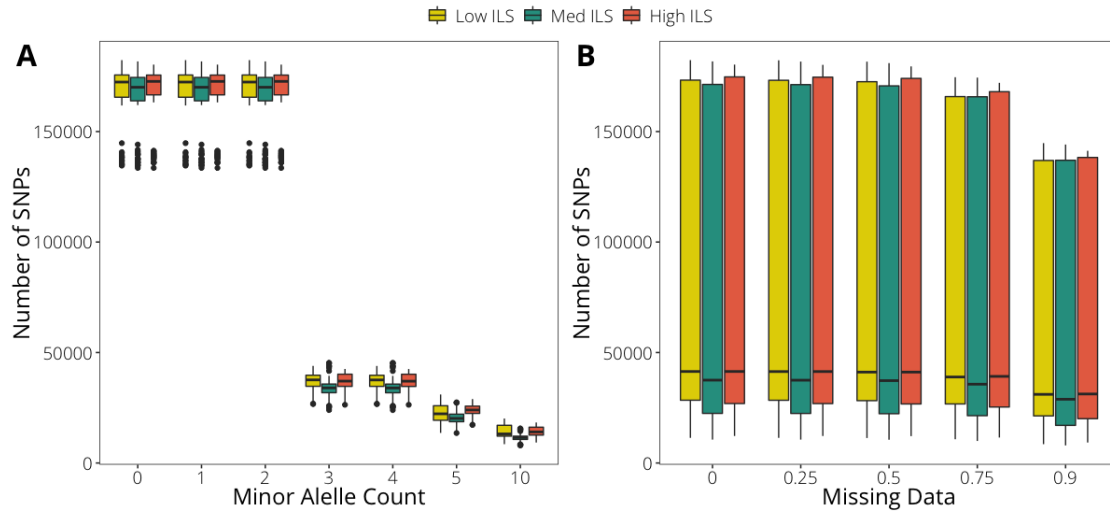


Fig. S2. Relationship within the full simulation data sets between (A) minor allele count threshold and the number of sites retained in a data set, and (B) missing data threshold and the number of sites retained in a data set, with colors indicating the tree height.

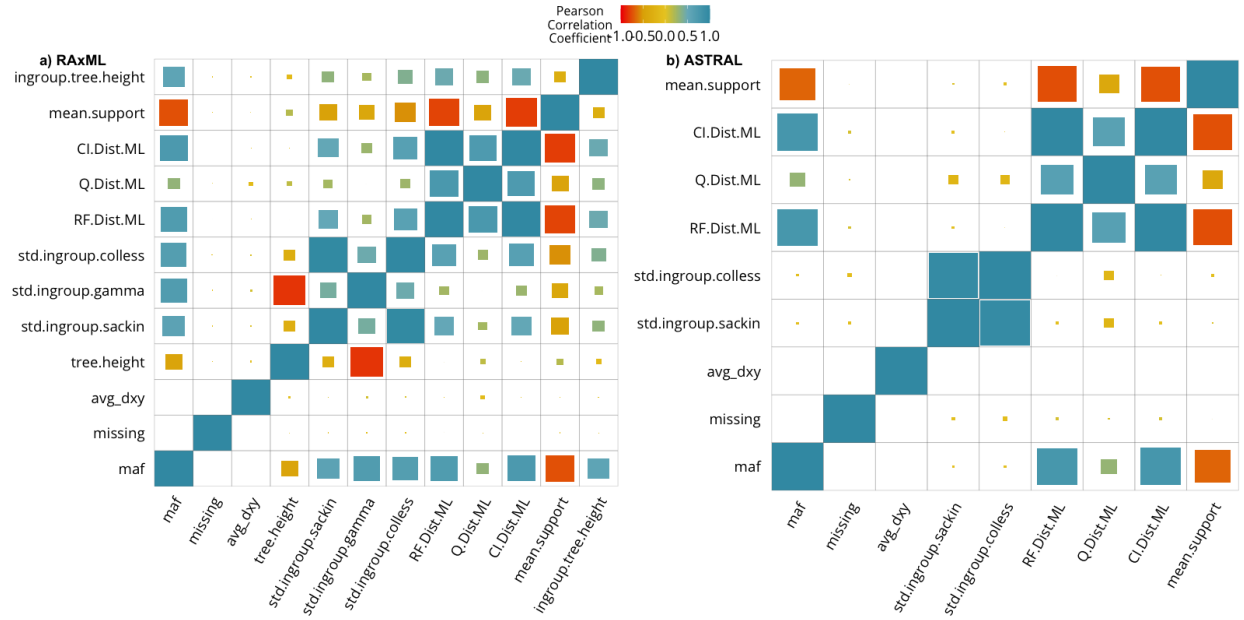


Fig. S3. Correlation matrix between output tree characteristics for the full simulation data sets, for phylogenies inferred in **RAxML** (a) and **ASTRAL** (b). Color and size correspond to the magnitude of the Pearson's correlation between the two parameters. Variables are overall tree height (tree.height), standardized Sackin imbalance of the ingroup (std.ingroup.sackin), standardized γ of the ingroup (std.ingroup.gamma), standardized Colless imbalance of the ingroup (std.ingroup.colless), Robinson Foulds distance to the true tree (RF.Dist.ML), quartet distance to the true tree (Q.Dist.ML), mean bootstrap support of all nodes in the tree (mean.support), height of the ingroup (ingroup.tree.height), and clustering information distance to the true tree (CI.Dist.ML). Statistics requiring branch lengths were not calculated for **ASTRAL**-inferred trees.

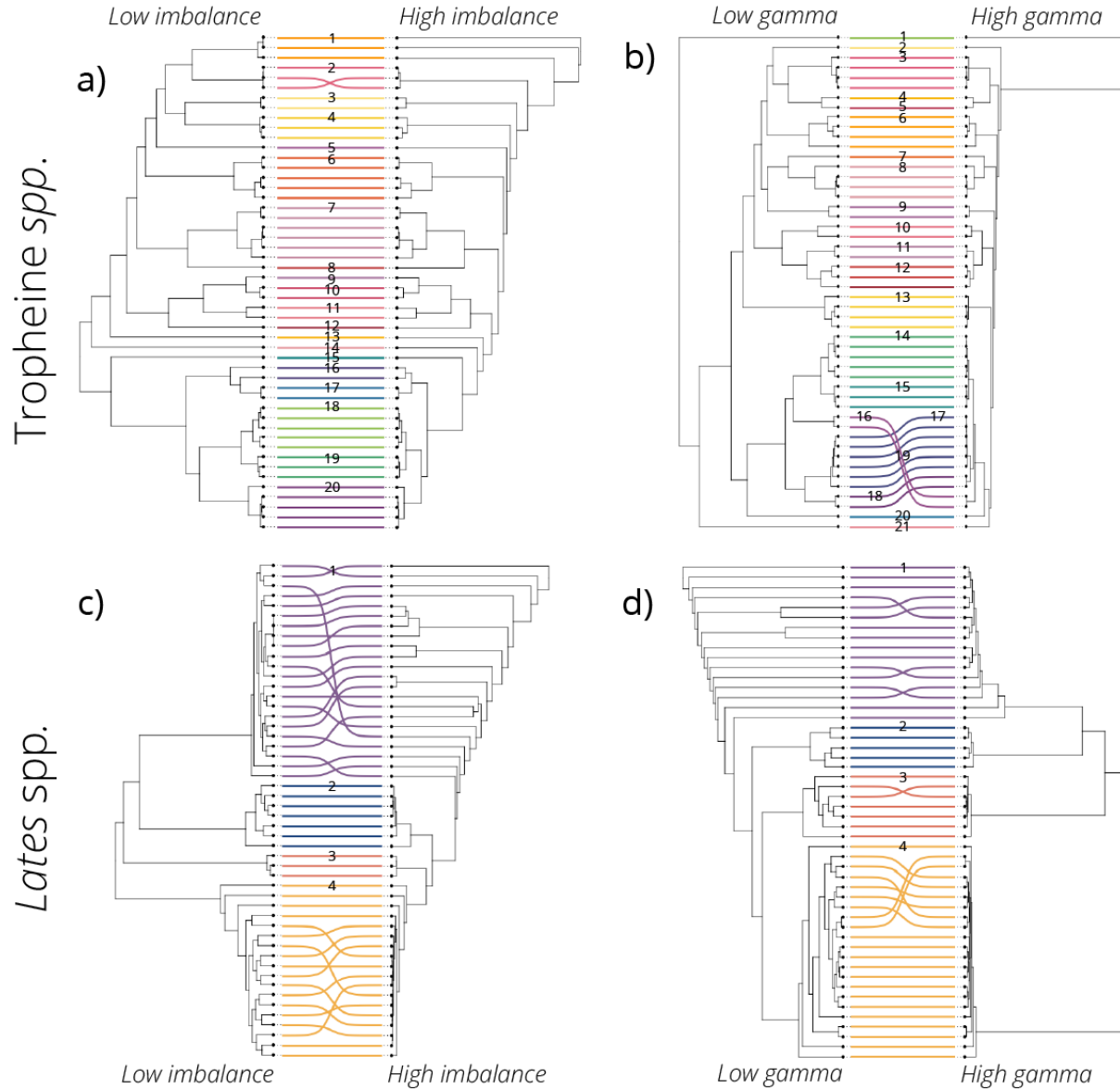


Fig. S4. Examples of inferred trees for tropheine (top) and *Lates* datasets, demonstrating extreme values for Colless' imbalance (left) and gamma (right). In each comparison, the two paired trees have the same individuals and differ only in filtering parameters. Each individual is connected to itself by a line colored according to species identity. For the tropheine phylogenies in (a), species identities are 1: *Tropheus annectens*, 2: *Tropheus kaiser*, 3: *Tropheus* sp. "kirschfleck", 4: *Tropheus* sp. "crescentic", 5: *Tropheus brichardi*, 6: *Tropheus duboisi*, 7: *Pundamilia nyererei*, 8: *Lobochilotes labiatus*, 9: *Petrochromis* sp. "green", 10: *Petrochromis moshi*, 11: *Petrochromis* sp. "kazumbe", 12: *Petrochromis* cf. *polyodon*, 13: *Petrochromis fasciolatus*, 14: *Petrochromis orthognathus*, 15: *Petrochromis famula*, 16: *Simochromis diagramma*, 17: *Limnotilapia dardennii*, 18: *Ctenochromis horei*, 19: *Pseudosimochromis marginatus*, 20: *Pseudosimochromis babaulti*. For the tropheine phylogenies in (b), species are 1: *Xenotilapia sima*, 2: *Limnotilapia dardennii*, 3: *Pseudosimochromis marginatus*, 4: *Pseudosimochromis babaulti*, 5: *Pseudosimochromis margaritae*, 6: *Petrochromis famula*, 7: *Petrochromis fasciolatus*, 8: *Petrochromis orthognathus*, 9: *Tropheus kaiser*, 10: *Tropheus* sp. "kirschfleck", 11: *Lobochilotes labiatus*, 12: *Petrochromis moshi*, 13: *Petrochromis* sp. "kazumbe", 14: *Petrochromis* cf. *polyodon*, 15: *Simochromis diagramma*, 16: *Tropheus annectens*, 17: *Tropheus* sp. "crescentic", 18: *Tropheus brichardi*, 19: *Tropheus duboisi*, 20: *Pundamilia nyererei*. In both (c) and (d), the *Lates* species identities are 1: *L. mariae*, 2: *L. angustifrons*, 3: *L. microlepis*, 4: *L. stappersii*.

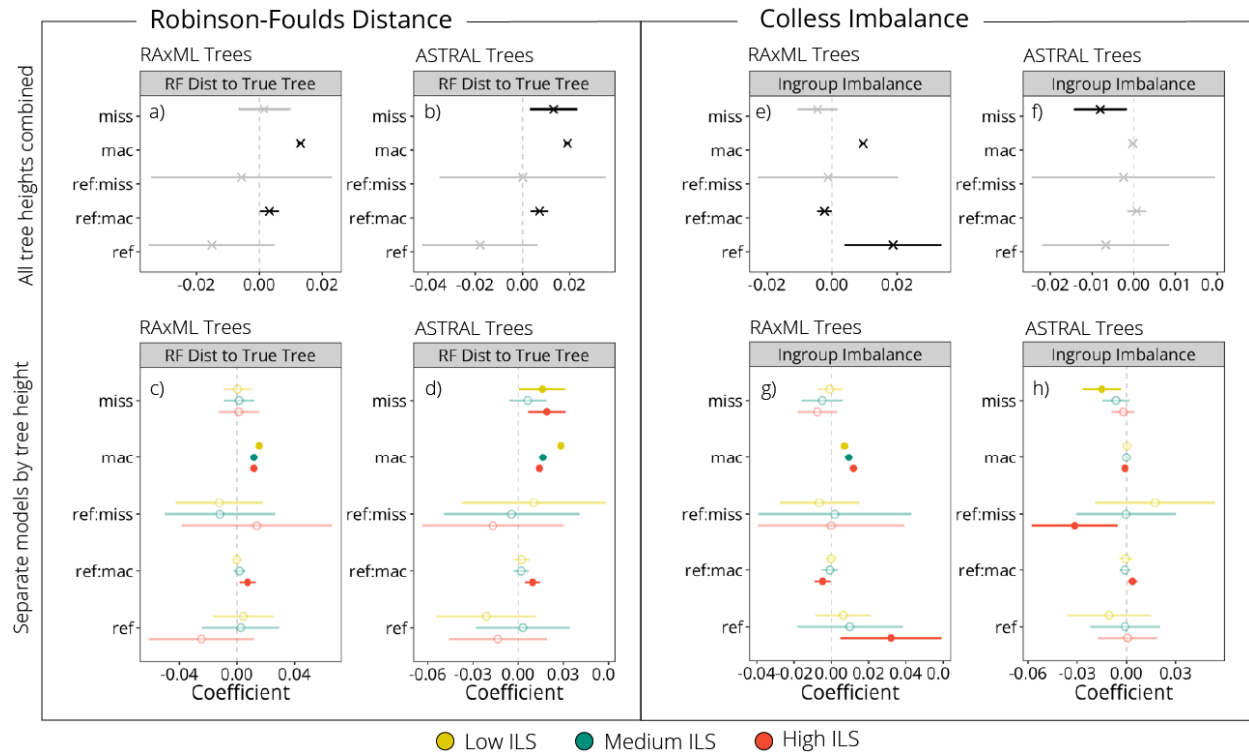


Fig. S5. Minor allele count (MAC) threshold had the greatest effect on the Robinson-Foulds distance (RF Dist) from an output tree to the true tree for both **RAxML** and **ASTRAL** output trees, with some interaction between minor allele count threshold and distance to the reference genome and tree height (i.e., amount of incomplete lineage sorting in the true tree). For predicting tree imbalance, MAC threshold was the strongest predictor for **RAxML** trees, while imbalance for **ASTRAL** trees was more dependent on the missing data cutoff. Plots in both (a)–(d) show coefficient estimates and 95% confidence intervals for each predictor variable from linear mixed models with Robinson-Foulds distance to the true tree as the response variable, while (e)–(h) are coefficient estimates and 95% confidence intervals for each predictor variable from linear mixed models with Colless' imbalance statistic as the response variable. Filled in and darkened symbols indicate significant positive or negative relationships, while faded symbols are predictors with confidence intervals overlapping zero. Plots in (a),(b), (e), and (f) are the results for all tree types combined, while those in (c), (d), (g), and (h) were modeled individually for each of the three simulated ILS levels. For model variables, mac = minor allele count threshold, avg_dxy = mean distance from ingroup taxa to the reference genome, miss = maximum missing data allowed, and colons show interaction terms between the given variables.

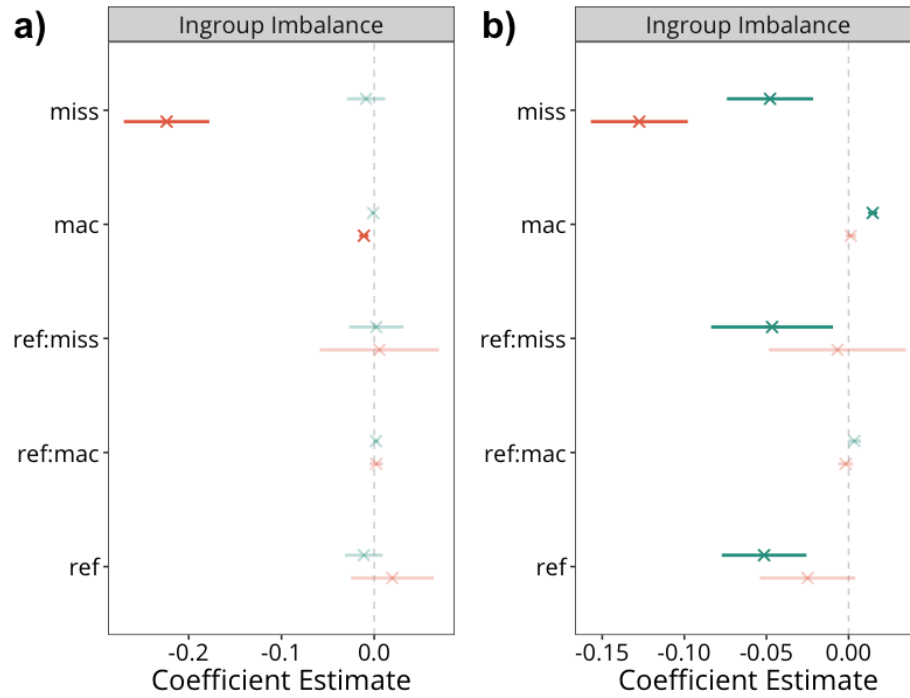


Fig. S6. A comparison of model coefficients for trees inferred using (a) **RAxML** and (b) **ASTRAL** for the Tropheine (orange) and *Lates* (teal) datasets. For the Tropheine dataset, missing data continued to be the strongest predictor of imbalance for the **ASTRAL** trees, minor allele count (MAC) threshold was a stronger predictor of imbalance, and the interaction between the reference genome and the missing data cutoff was more important in the **ASTRAL** trees than those inferred using **RAxML**. For the *Lates* dataset, missing data and MAC thresholds were both stronger predictors of imbalance in the **ASTRAL** trees, and the reference genome also had a stronger effect. Plots show coefficient estimates and 95% confidence intervals for each predictor variable from linear mixed models with the Colless' imbalance statistic of the ingroup as the response variable, with missing data cutoff (miss), minor allele count (mac), reference genome (ref), the interaction between missing data and reference genome choice (int:miss), and the interaction between minor allele count and reference genome choice (int:mac) as explanatory variables. Filled in and darkened symbols indicate significant positive or negative relationships, while faded symbols are predictors with confidence intervals overlapping zero.

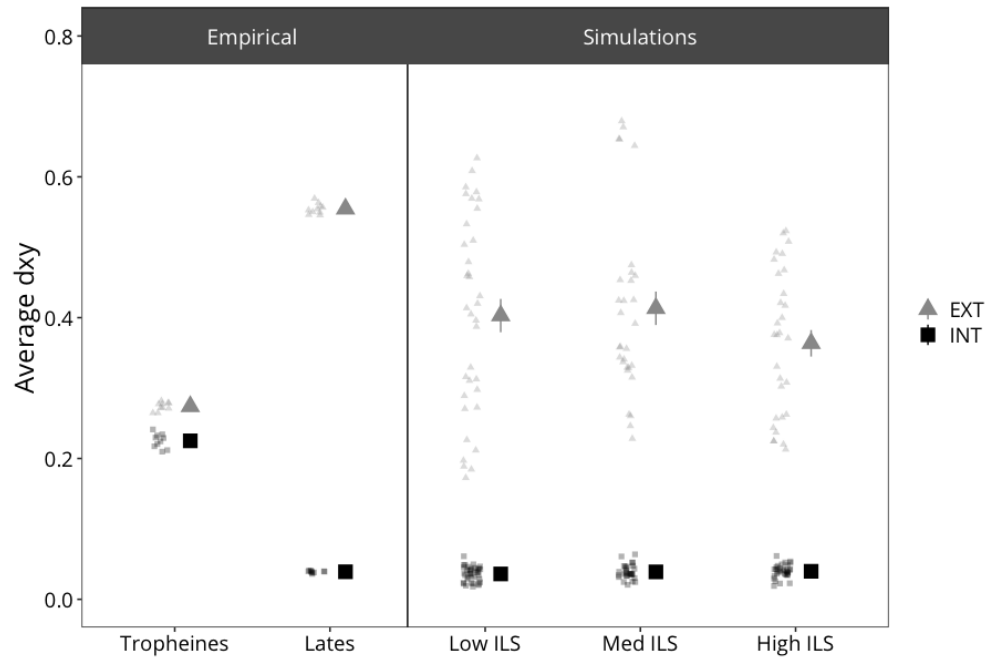


Fig. S7. A comparison of the average distance of ingroup taxa to the reference genome for the empirical (left) and simulated (right) datasets. Each light-colored point represents one replicate dataset, and larger dark points show the mean and standard error for each category.

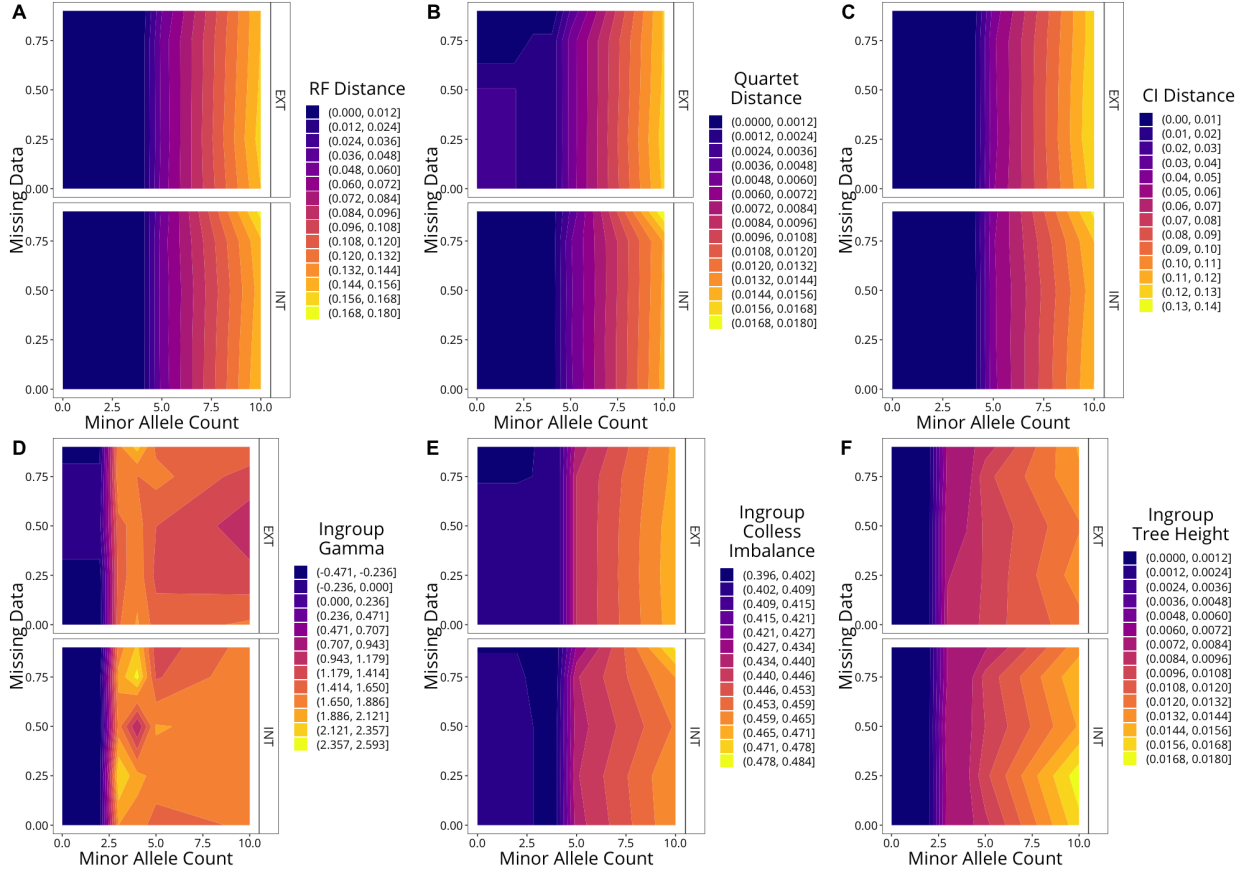


Fig. S8. Heatmaps for the **low ILS** trees indicating how each of our tree descriptive parameters varies with minor allele count and missing data thresholds, with darker colors indicating lower values and lighter colors indicating higher values. Three of the parameters, Robinson-Foulds (RF) distance, quartet distance, and clustering information (CI) distance, are measures of topological dissimilarity compared to the true simulated tree, and lower values indicate topologies closer to the true tree. The gamma and Colless imbalance statistics are measures of tree center of gravity and tree shape, with means for the true trees around $I_c=0.4$ and $\gamma=0$.

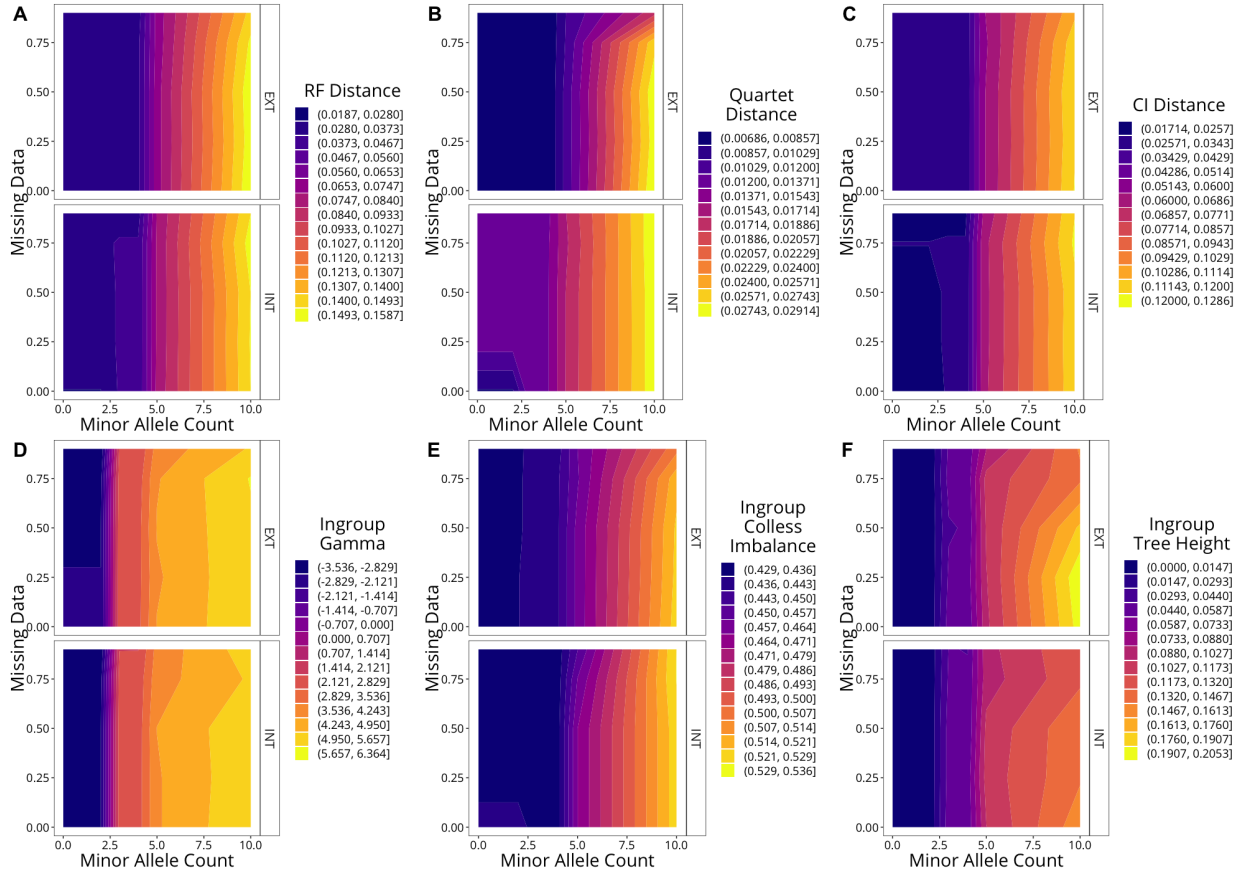


Fig. S9. Heatmaps for the **medium ILS** trees indicating how each of our tree descriptive parameters varies in two-dimensional space with minor allele count and missing data thresholds, with darker colors indicating lower values and lighter colors indicating higher values. Three of the parameters, Robinson-Foulds (RF) distance, quartet distance, and clustering information (CI) distance, are measures of topological dissimilarity compared to the true simulated tree, and lower values indicate topologies closer to the true tree. The gamma and Colless imbalance statistics are measures of tree center of gravity and tree shape, with means for the true trees around $I_c=0.4$ and $\gamma=0$.

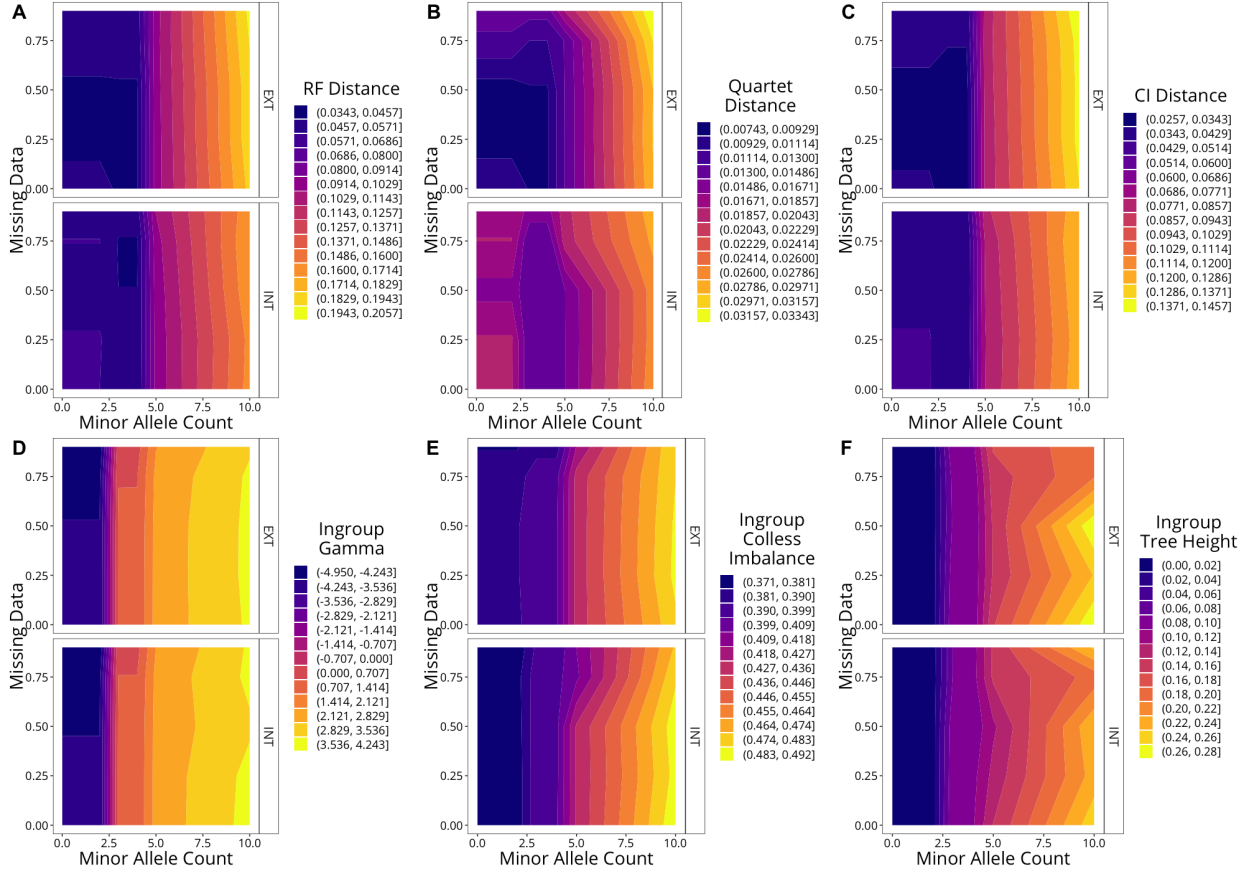


Fig. S10. Heatmaps for the **high ILS** trees indicating how each of our tree descriptive parameters varies in two-dimensional space with minor allele count and missing data thresholds, with darker colors indicating lower values and lighter colors indicating higher values. Three of the parameters, Robinson-Foulds (RF) distance, quartet distance, and clustering information (CI) distance, are measures of topological dissimilarity compared to the true simulated tree, and lower values indicate topologies closer to the true tree. The gamma and Colless imbalance statistics are measures of tree center of gravity and tree shape, with means for the true trees around $I_c=0.4$ and $\gamma=0$.

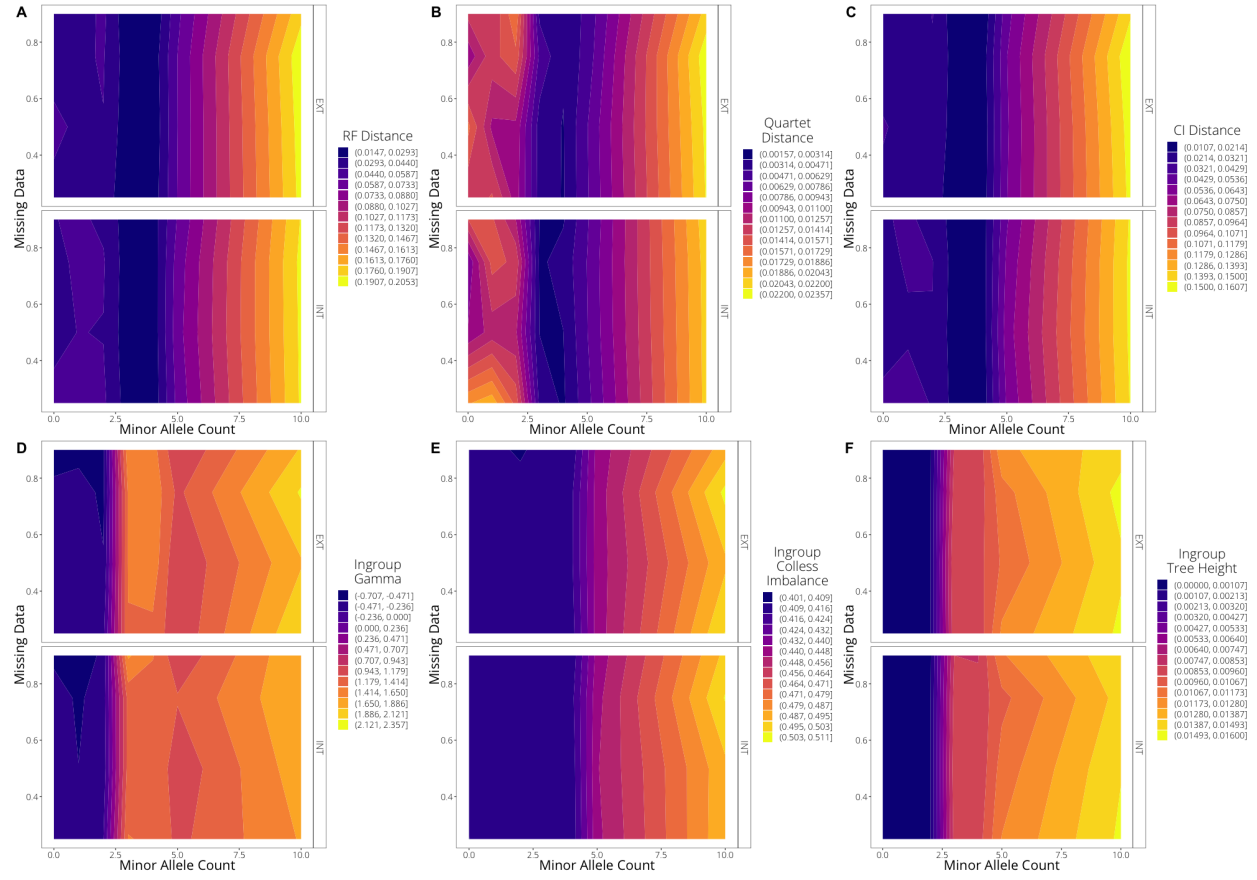


Fig. S11. Heatmaps for the **subsampled low ILS** trees indicating how each of our tree descriptive parameters varies with minor allele count and missing data thresholds, with darker colors indicating lower values and lighter colors indicating higher values. Three of the parameters, Robinson-Foulds (RF) distance, quartet distance, and clustering information (CI) distance, are measures of topological dissimilarity compared to the true simulated tree, and lower values indicate topologies closer to the true tree. The gamma and Colless imbalance statistics are measures of tree center of gravity and tree shape, with means for the true trees around $I_c=0.4$ and $\gamma=0$.

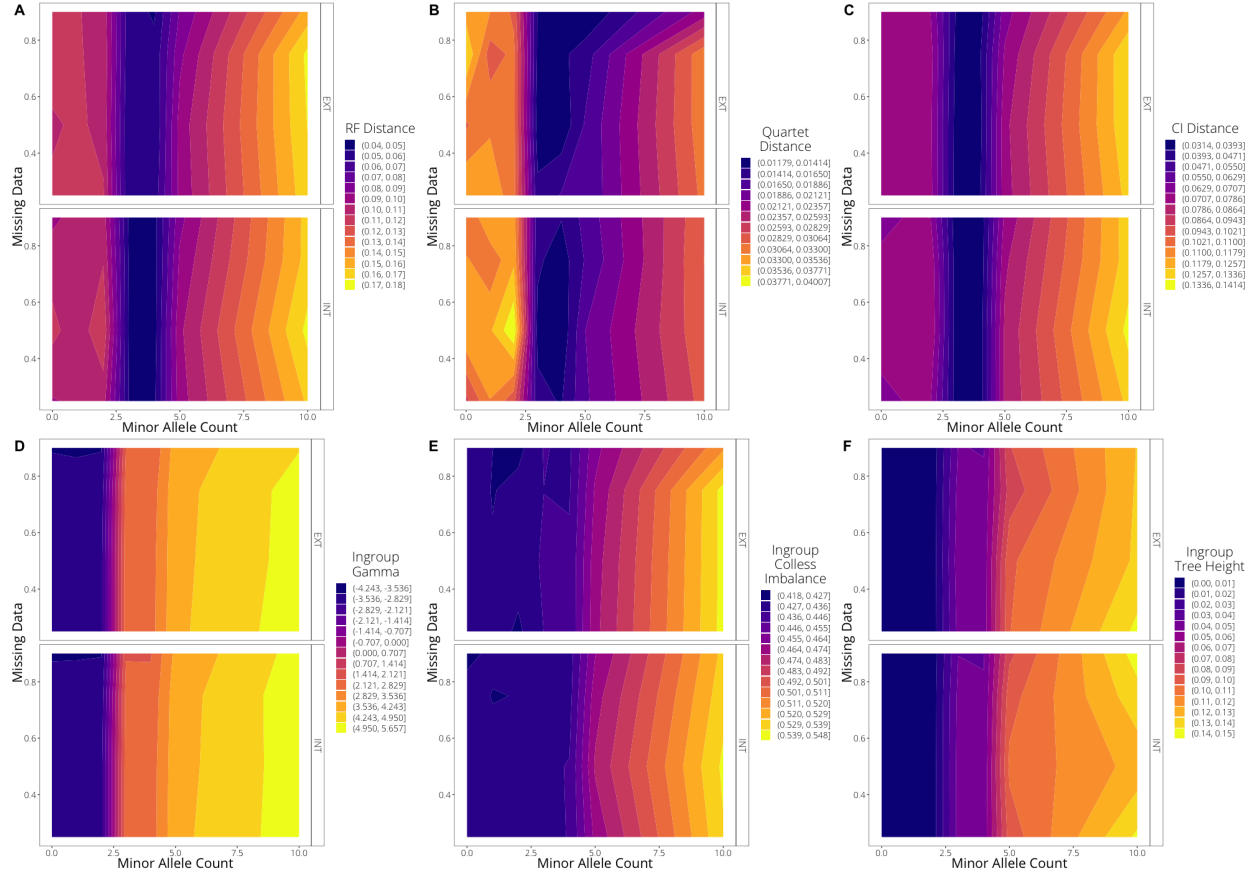


Fig. S12. Heatmaps for the **subsampled medium ILS trees** indicating how each of our tree descriptive parameters varies in two-dimensional space with minor allele count and missing data thresholds, with darker colors indicating lower values and lighter colors indicating higher values. Three of the parameters, Robinson-Foulds (RF) distance, quartet distance, and clustering information (CI) distance, are measures of topological dissimilarity compared to the true simulated tree, and lower values indicate topologies closer to the true tree. The gamma and Colless imbalance statistics are measures of tree center of gravity and tree shape, with means for the true trees around $I_c=0.4$ and $\gamma=0$.

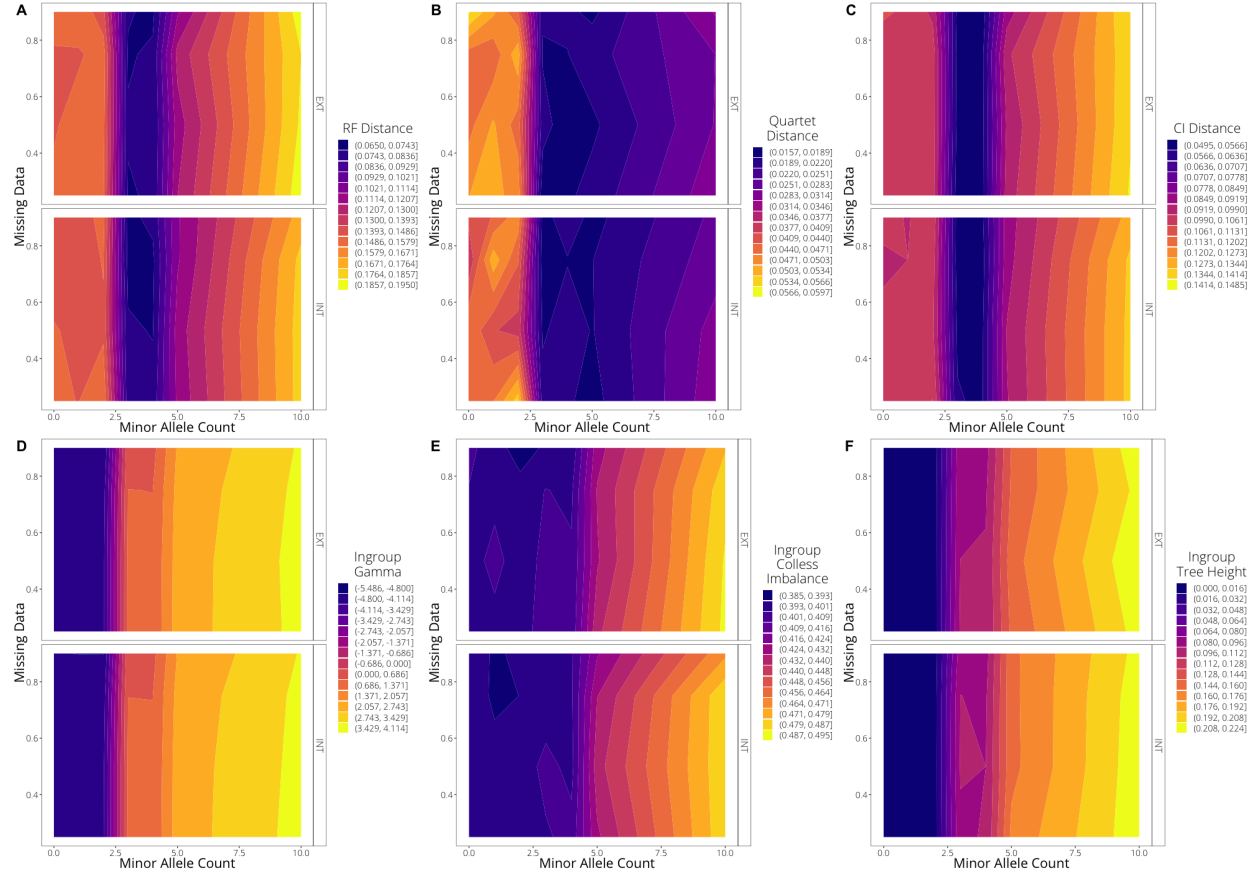


Fig. S13. Heatmaps for the **subsampled high ILS trees** indicating how each of our tree descriptive parameters varies in two-dimensional space with minor allele count and missing data thresholds, with darker colors indicating lower values and lighter colors indicating higher values. Three of the parameters, Robinson-Foulds (RF) distance, quartet distance, and clustering information (CI) distance, are measures of topological dissimilarity compared to the true simulated tree, and lower values indicate topologies closer to the true tree. The gamma and Colless imbalance statistics are measures of tree center of gravity and tree shape, with means for the true trees around $I_c=0.4$ and $\gamma=0$.

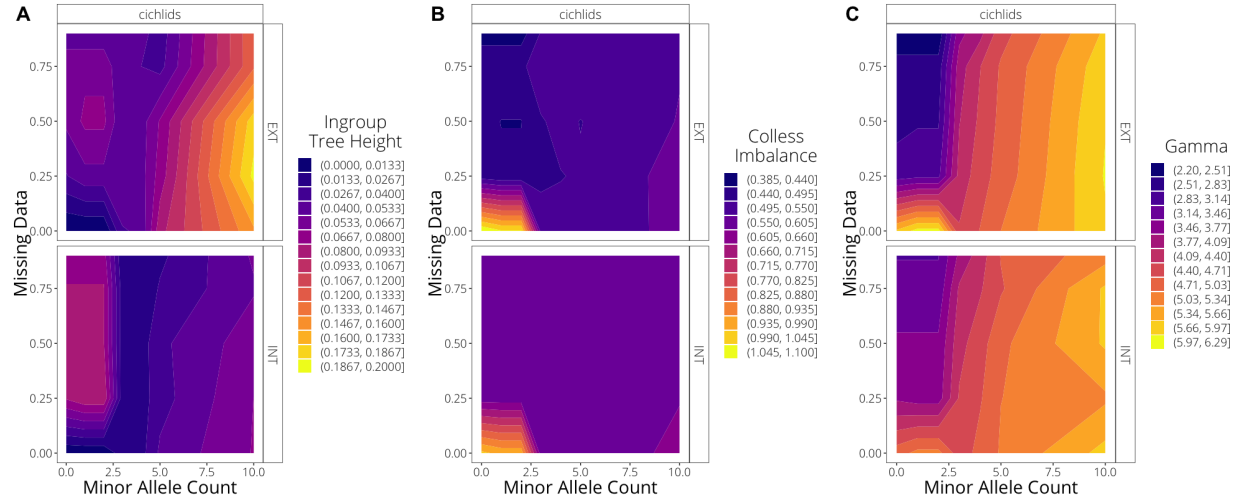


Fig. S14. Heatmaps for the **empirical Tropheine** trees indicating how each of our tree descriptive parameters varies in two-dimensional space with minor allele count and missing data thresholds, with darker colors indicating lower values and lighter colors indicating higher values. The gamma and Colless imbalance statistics are measures of tree center of gravity and tree shape, respectively.

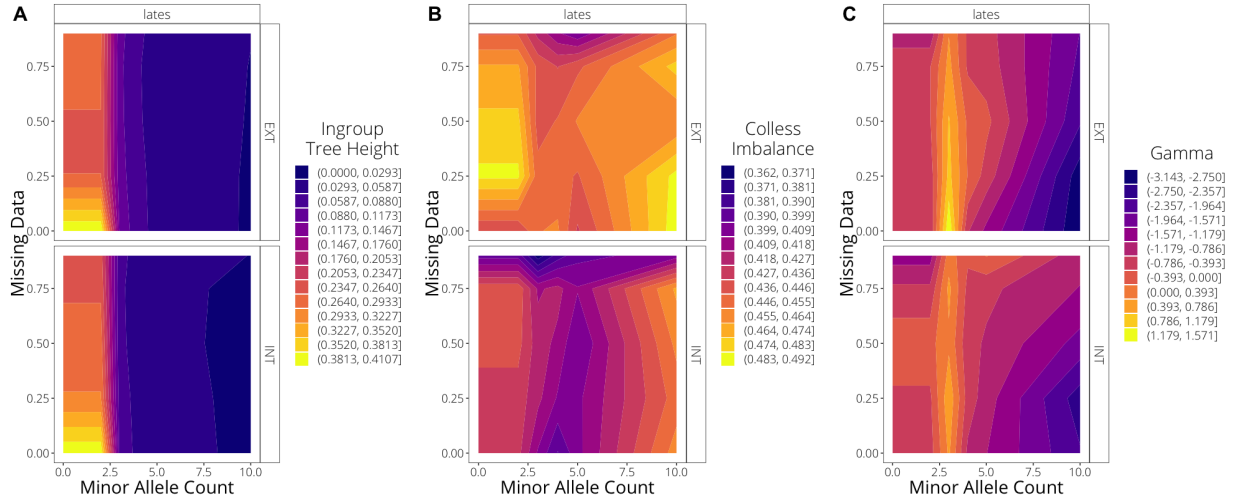


Fig. S15. Heatmaps for the **empirical *Lates* trees** indicating how each of our tree descriptive parameters varies with minor allele count and missing data thresholds, with darker colors indicating lower values and lighter colors indicating higher values. The gamma and Colless imbalance statistics are measures of tree center of gravity and tree shape, respectively.

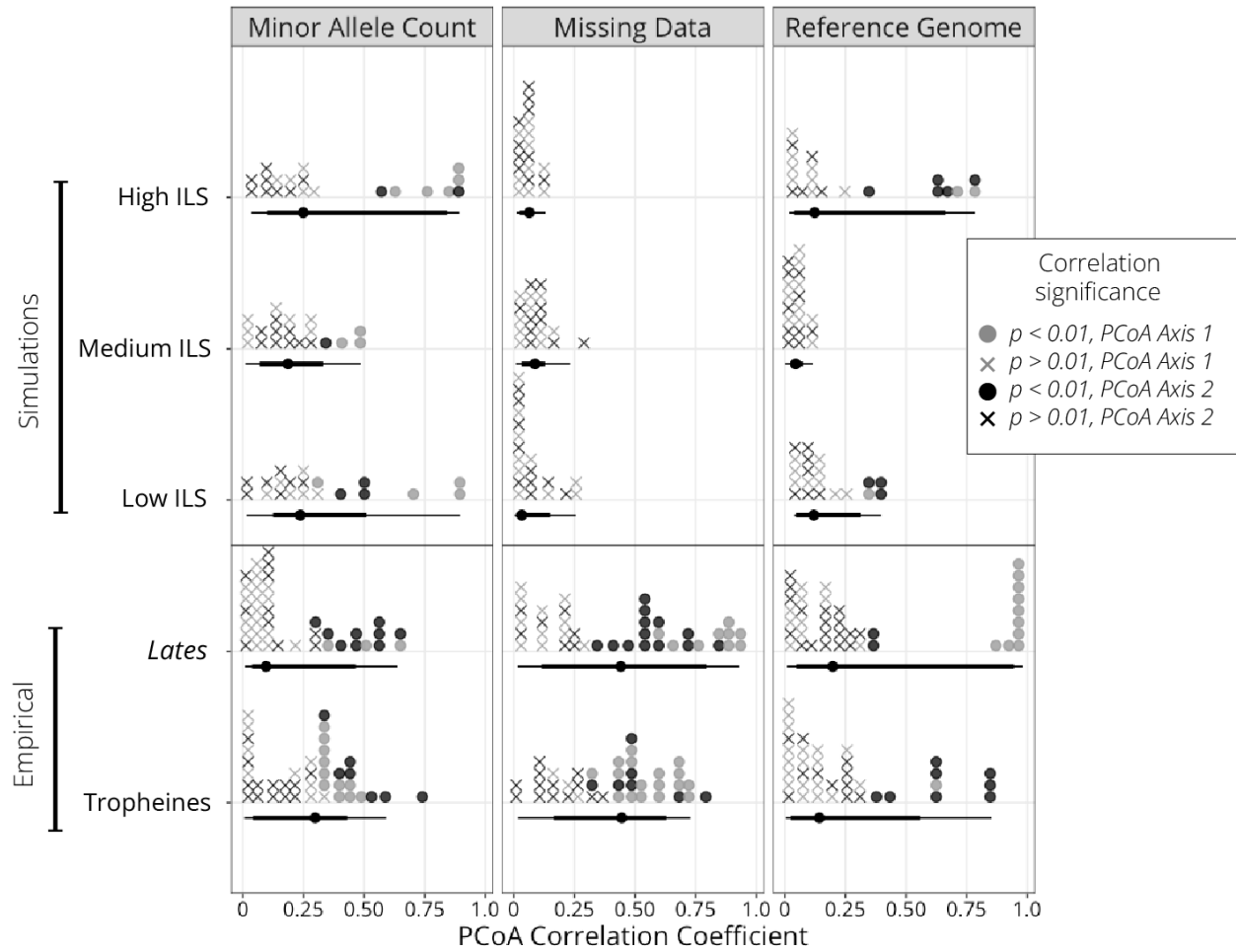


Fig. S16. Principal coordinate analyses (PCoAs) based on pairwise Robinson-Foulds distances between ASTRAL-inferred trees demonstrate the relationship between simulated and true trees in two dimensional tree space. Plot shows correlation coefficients between each of the three bioinformatic choices and the first two PCoA axes for all simulated and empirical iterations, demonstrating the relative importance of minor allele count in the simulated data and missing data in the empirical data sets, while reference genome was important for all. Dotplots are colored by correlation significance (p -value < 0.01 in solid circles, p -value ≥ 0.01 shown with x's), assessed using Pearson's correlation coefficients. Lines below dots show median, 95% quantiles (thick lines), and full data range (thin lines) for each group of analyses.

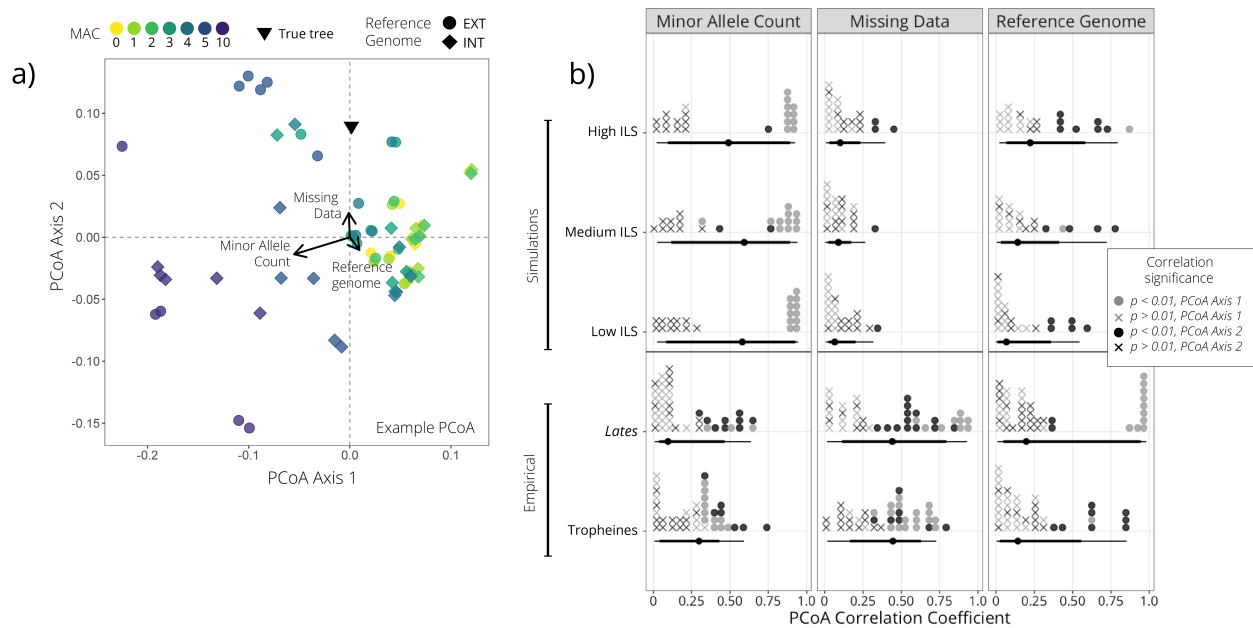


Fig. S17. Principal coordinate analyses (PCoAs) based on pairwise Robinson-Foulds distances between ASTRAL-inferred trees demonstrate the relationship between simulated and true trees in two dimensional tree space. (a) Example PCoA plot of trees from one iteration of simulated data (high ILS, simulation 18), with the true tree indicated and correlations between PCoA axes and each of reference genome choice, minor allele count (MAC), and missing data threshold visualized via arrows. (b) Summary of correlations between each of the three bioinformatic choices and the first two PCoA axes for all simulated and empirical iterations, demonstrating the relative importance of minor allele count in the simulated data and missing data in the empirical data sets, while reference genome was important for all. Dotplots show the distribution of values, colored by correlation significance (p-value < 0.01 in solid circles, p-value > 0.01 shown with x's), assessed using Pearson's correlation coefficients. Lines below dots show median, 95% quantiles (thick lines), and full data range (thin lines) for each group of analyses.

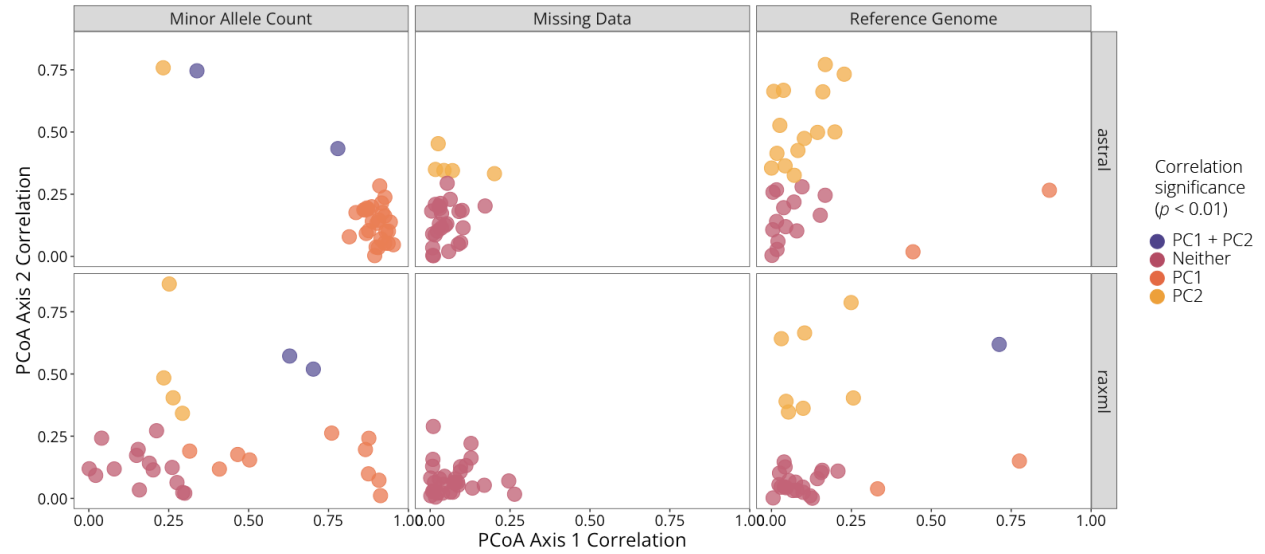


Fig. S18. Comparison of PCoA loadings between **RAxML** and **ASTRAL** trees, based on all pairwise Robinson-Foulds distances between trees inferred with the given method. For each parameter, the correlation coefficient between the parameter and a tree's location in PCoA space for each dataset is plotted, with points colored by significance of the correlation (orange, PC1 $p < 0.01$; yellow, PC2 $p < 0.01$; purple, PC1 & PC2 both $p < 0.01$; pink, both $p \geq 0.01$). Using both phylogenetic inference methods, minor allele count and reference genome were generally strongly correlated with the first two principal component axes.

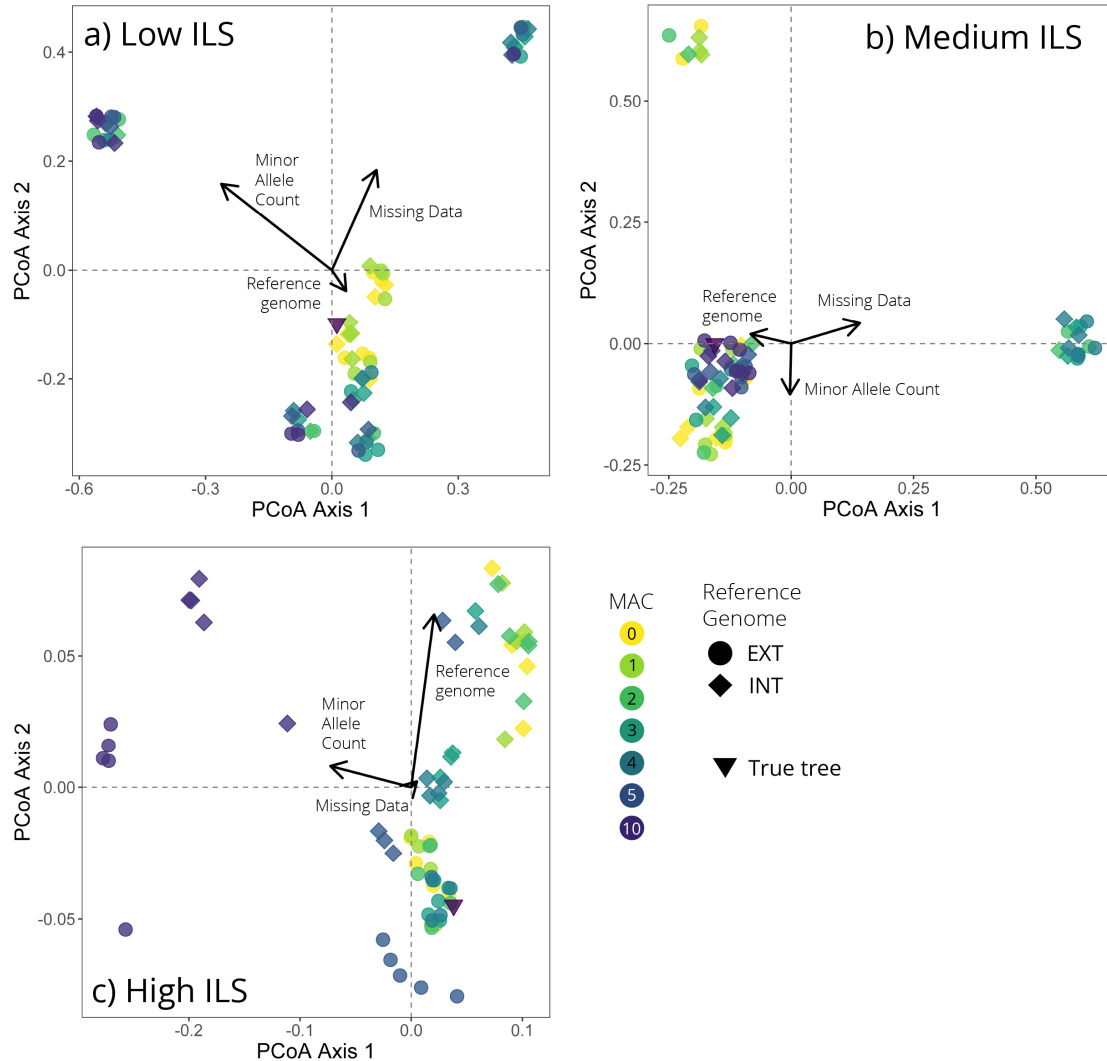


Fig. S19. Example plots of principal coordinates analysis (PCoA) visualizing distance between trees in tree space for (A) low ILS, (B), medium ILS, and (C) high ILS trees. For each plot, the true, simulated species tree is shown as a black triangle, while the other points are colored according to their minor allele count threshold and shape indicates the reference genome choice (circle = EXT/outgroup reference; diamond = INT/ingroup reference). Arrows on the plots show correlations of each of the three bioinformatic choices of interest with the first two PCoA axes, with the length of the arrow corresponding to the magnitude of correlation with the given bioinformatic choice. Note that points have been jittered slightly to make it clear when multiple trees overlap in tree space.

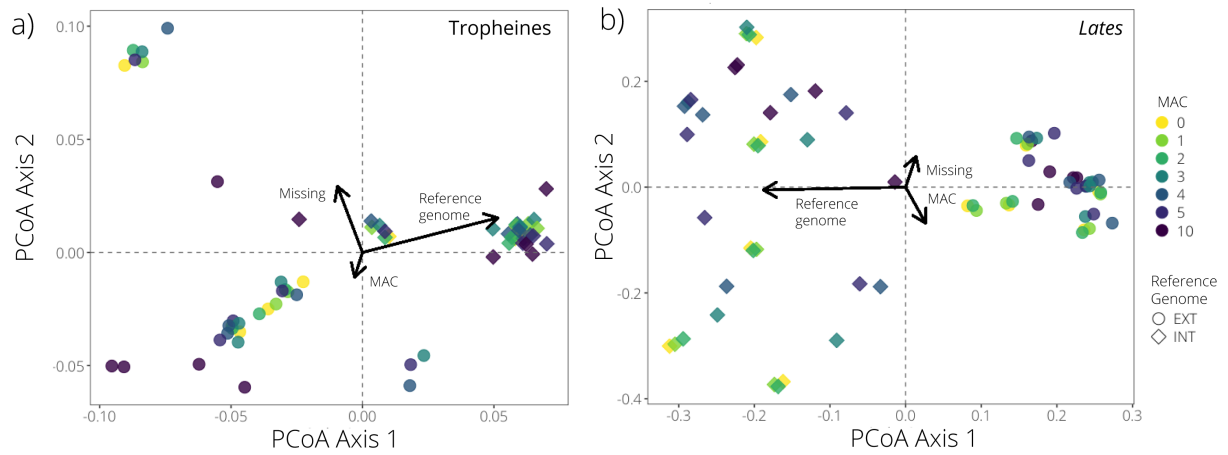


Fig. S20. Example plots of principal coordinates analysis (PCoA) visualizing distance between trees in tree space for (a) Tropheine and (b) *Lates* trees. For each plot, the points are colored according to their minor allele count threshold and shape indicates the reference genome choice (circle = EXT/outgroup reference; diamond = INT/ingroup reference). Arrows on the plots show correlations of each of the three bioinformatic choices of interest with the first two PCoA axes, with the length of the arrow corresponding to the magnitude of correlation with the given bioinformatic choice.

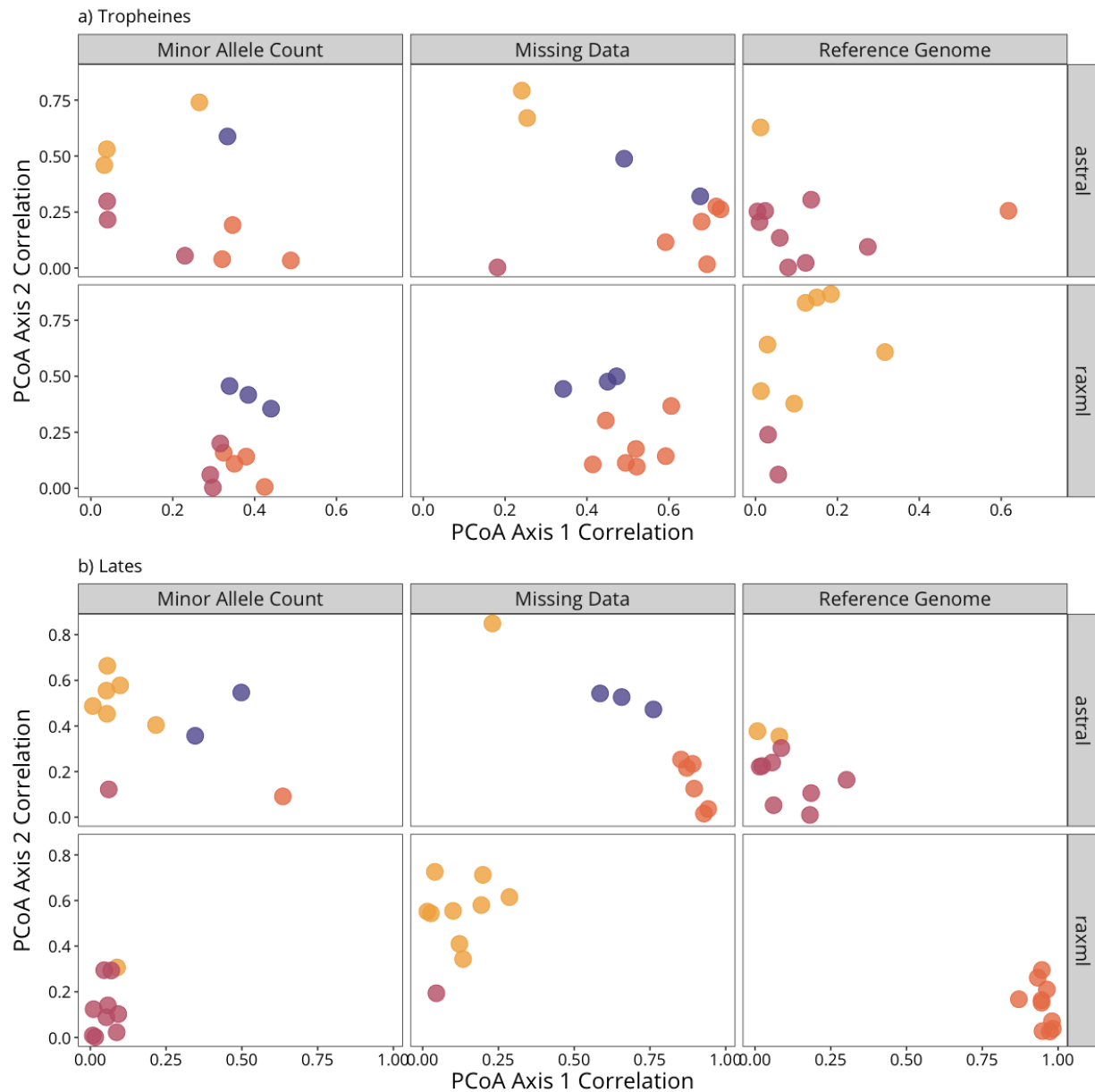


Fig. S21. Comparison of PCoA axis correlations between RAxML and ASTRAL trees for the empirical Tropheine (a) and *Lates* (b) datasets, based on all pairwise Robinson-Foulds distances between trees. For each parameter, the correlation coefficient between the parameter values and each tree's location in PCoA space for each dataset is plotted (i.e., each point is one of the ten replicate datasets), with points colored by significance of the correlation (orange, PC1 $p < 0.01$; yellow, PC2 $p < 0.01$; purple, PC1 and PC2 both $p < 0.01$; pink, both $p \geq 0.01$).

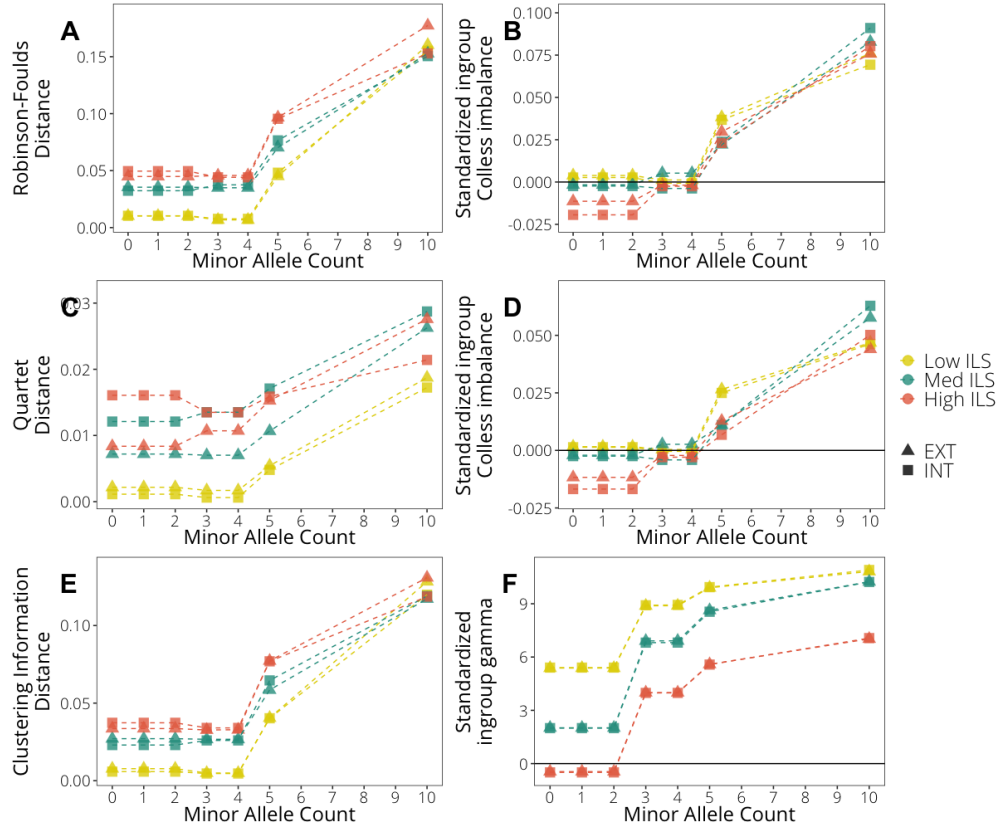


Fig. S22. Patterns of variation for the full simulation data sets in the three metrics of distance to the true tree, two metrics of tree imbalance, and tree center of gravity (γ) across minor allele counts, with shapes indicating the reference used (EXT, outgroup reference; INT, ingroup reference) and colors indicating the level of ILS. For topological accuracy metrics (Robinson Foulds, quartet, and clustering information distances to the true tree), higher values indicate topologies more biased away from the true tree, and all have been normalized to be bounded by $0 \leq \text{distance} \leq 1$. Measures of Colless imbalance, Sackin imbalance, and γ have been standardized so that a value of 0 is equal to the measure in the true tree.

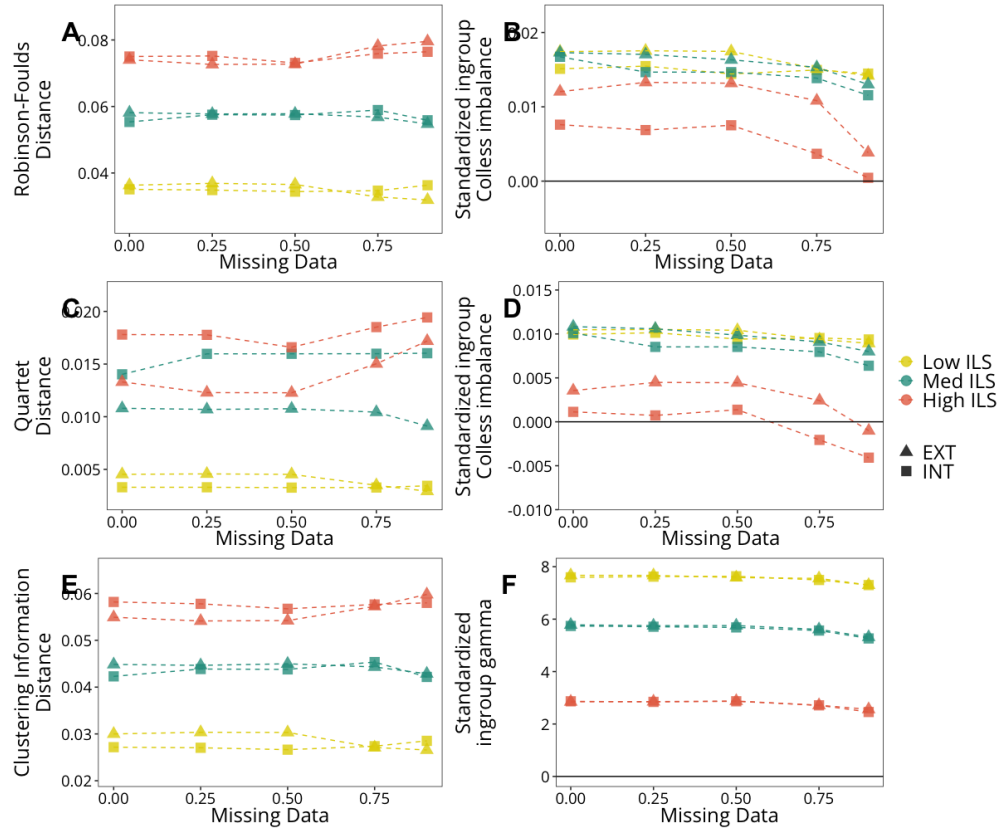


Fig. S23. Patterns of variation for the full simulation data sets in the three metrics of distance to the true tree, two metrics of tree imbalance, and tree center of gravity (γ) across missing data thresholds, with shapes indicating the reference used (EXT, outgroup reference; INT, ingroup reference) and colors indicating the level of ILS. For topological accuracy metrics (Robinson Foulds, quartet, and clustering information distances to the true tree), higher values indicate topologies more biased away from the true tree, and all have been normalized to be bounded by $0 \leq \text{distance} \leq 1$. Measures of Colless imbalance, Sackin imbalance, and γ have been standardized so that a value of 0 is equal to the measure in the true tree.

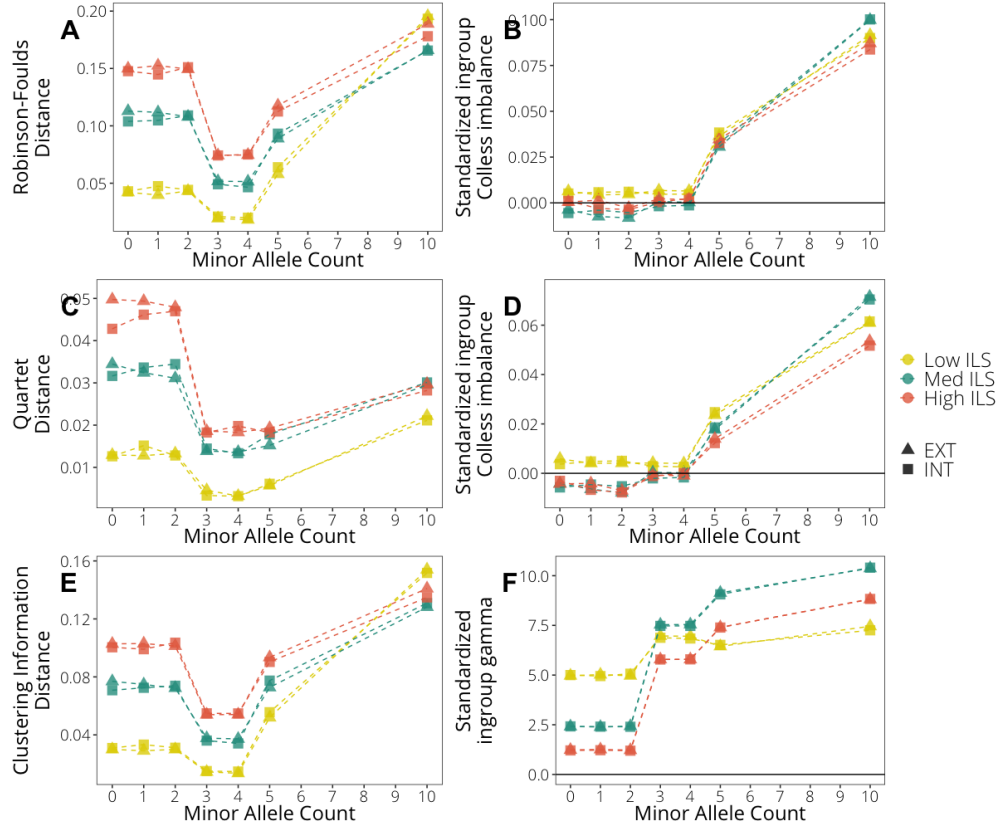


Fig. S24. Patterns of variation for the subsampled simulation data sets in the three metrics of distance to the true tree, two metrics of tree imbalance, and tree center of gravity (γ) across minor allele counts, with shapes indicating the reference used (EXT, outgroup reference; INT, ingroup reference) and colors indicating the level of ILS. For topological accuracy metrics (Robinson Foulds, quartet, and clustering information distances to the true tree), higher values indicate topologies more biased away from the true tree, and all have been normalized to be bounded by $0 \leq \text{distance} \leq 1$. Measures of Colless imbalance, Sackin imbalance, and γ have been standardized so that a value of 0 is equal to the measure in the true tree.

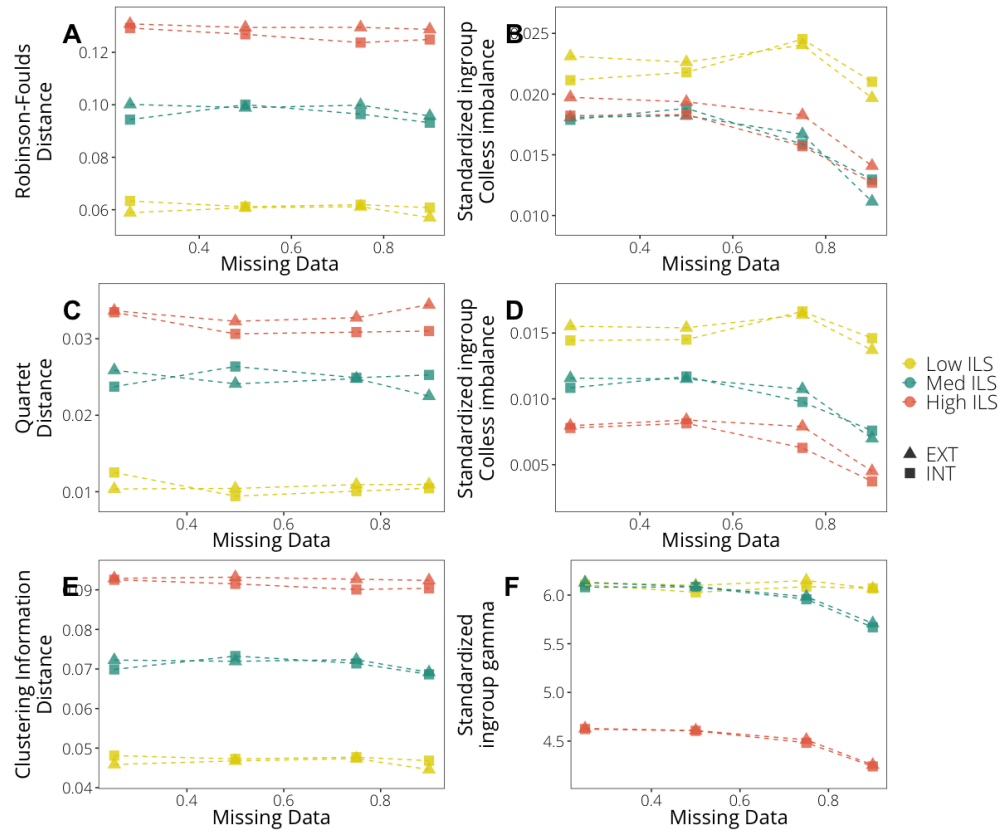


Fig. S25. Patterns of variation for the subsampled simulation data sets in the three metrics of distance to the true tree, two metrics of tree imbalance, and tree center of gravity (γ) across missing data thresholds, with shapes indicating the reference used (EXT, outgroup reference; INT, ingroup reference) and colors indicating the level of ILS. For topological accuracy metrics (Robinson Foulds, quartet, and clustering information distances to the true tree), higher values indicate topologies more biased away from the true tree, and all have been normalized to be bounded by $0 \leq \text{distance} \leq 1$. Measures of Colless imbalance, Sackin imbalance, and γ have been standardized so that a value of 0 is equal to the measure in the true tree.

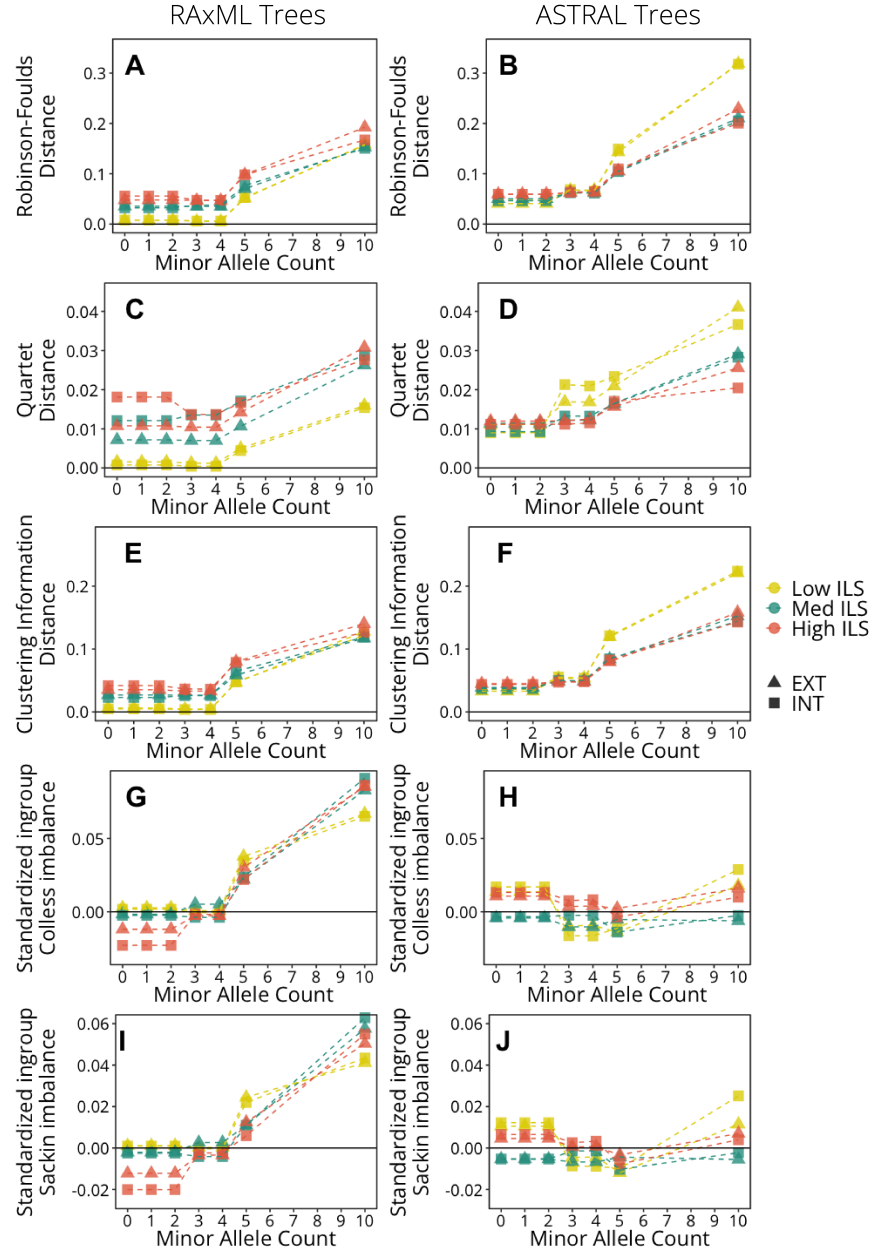


Fig. S26. Patterns of variation for the full simulation data sets for RAxML (left column) and ASTRAL (right column) trees in the three metrics of distance to the true tree and two metrics of tree imbalance across minor allele counts. Shapes indicate the reference that reads were aligned to (EXT, outgroup reference; INT, ingroup reference) and colors indicate the level of ILS in the true tree. For topological accuracy metrics (Robinson Foulds, quartet, and clustering information distances to the true tree), higher values indicate topologies more biased away from the true tree, and all have been normalized to be bounded by $0 \leq \text{distance} \leq 1$. Measures of Colless imbalance and Sackin imbalance have been standardized so that a value of 0 is equal to the measure in the true tree. In all plots, black lines indicate the value for the true tree.

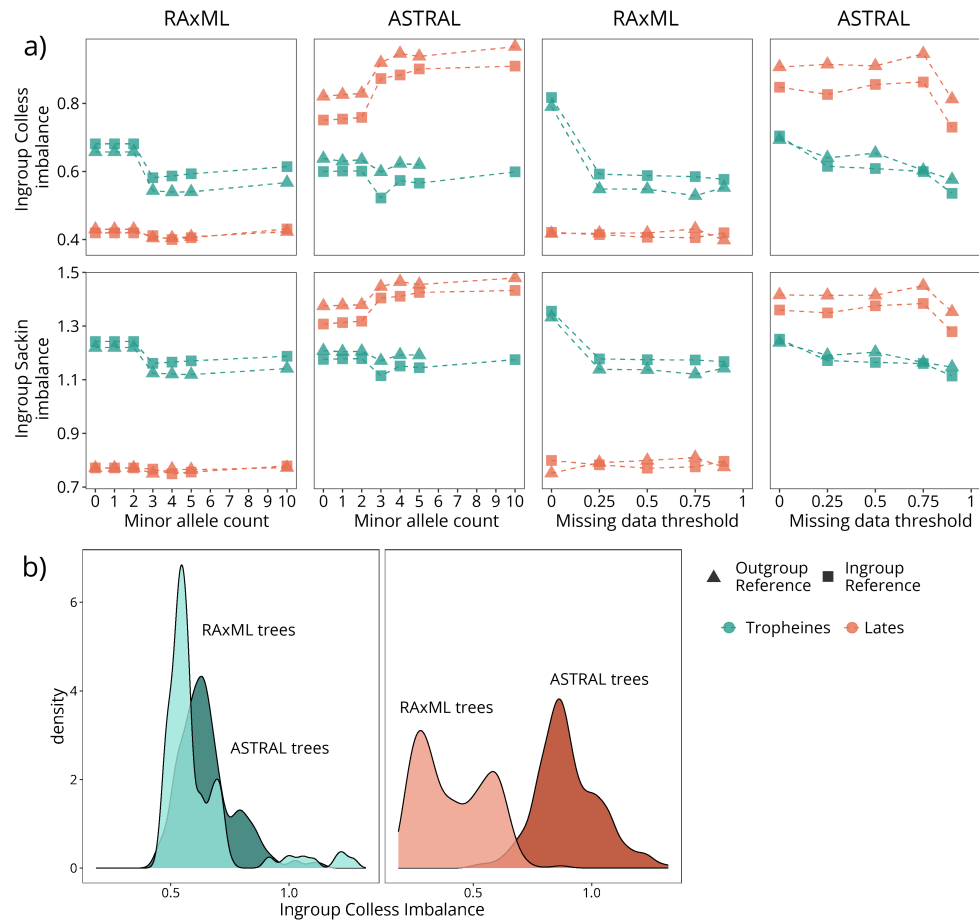


Fig. S27. Patterns of variation for the empirical data sets for RAxML (left column of each pair) and ASTRAL (right column of each pair) trees in the two metrics of tree imbalance across minor allele counts (left) and missing data (right). Shapes indicate the reference that reads were aligned to and colors indicate the data set. In (b), distributions of imbalance statistics are shown for each empirical data set comparing RAxML (light colors) to ASTRAL (dark colors) trees, demonstrating that trees inferred in ASTRAL are generally more imbalanced.

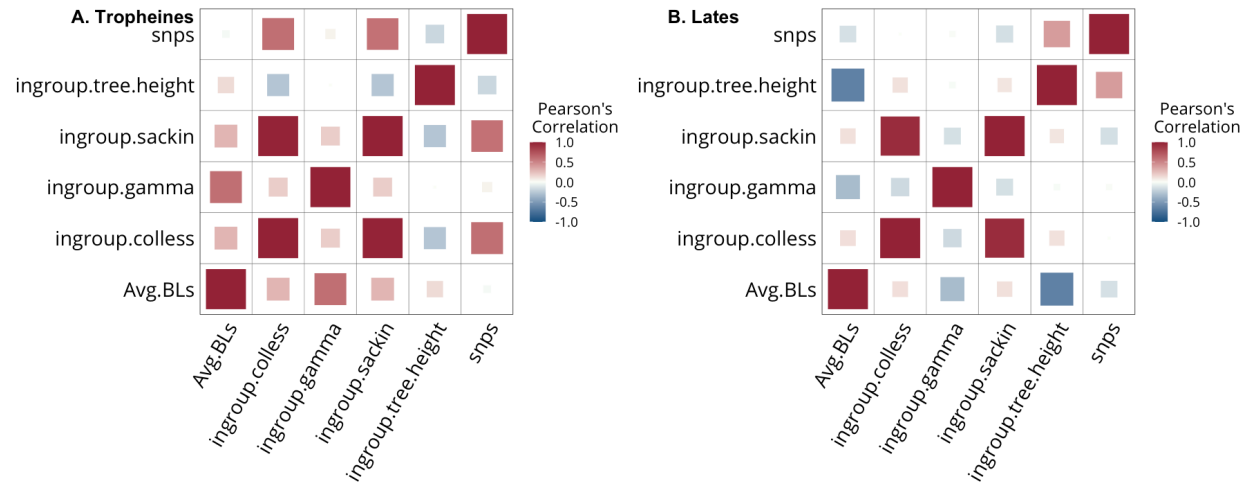


Fig. S28. Correlation matrix between output tree characteristics for the (A) *Tropheine* and (B) *Lates* empirical data sets. Color and size correspond to the magnitude of the correlation. Variables are ingroup tree height (ingroup.tree.height), Sackin imbalance of the ingroup (ingroup.sackin), γ of the ingroup (ingroup.gamma), Colless imbalance of the ingroup (ingroup.colless), and average branch lengths (Avg.BLs). White squares indicate pairs with non-significant Pearson's correlation coefficients at $\alpha = 0.05$.

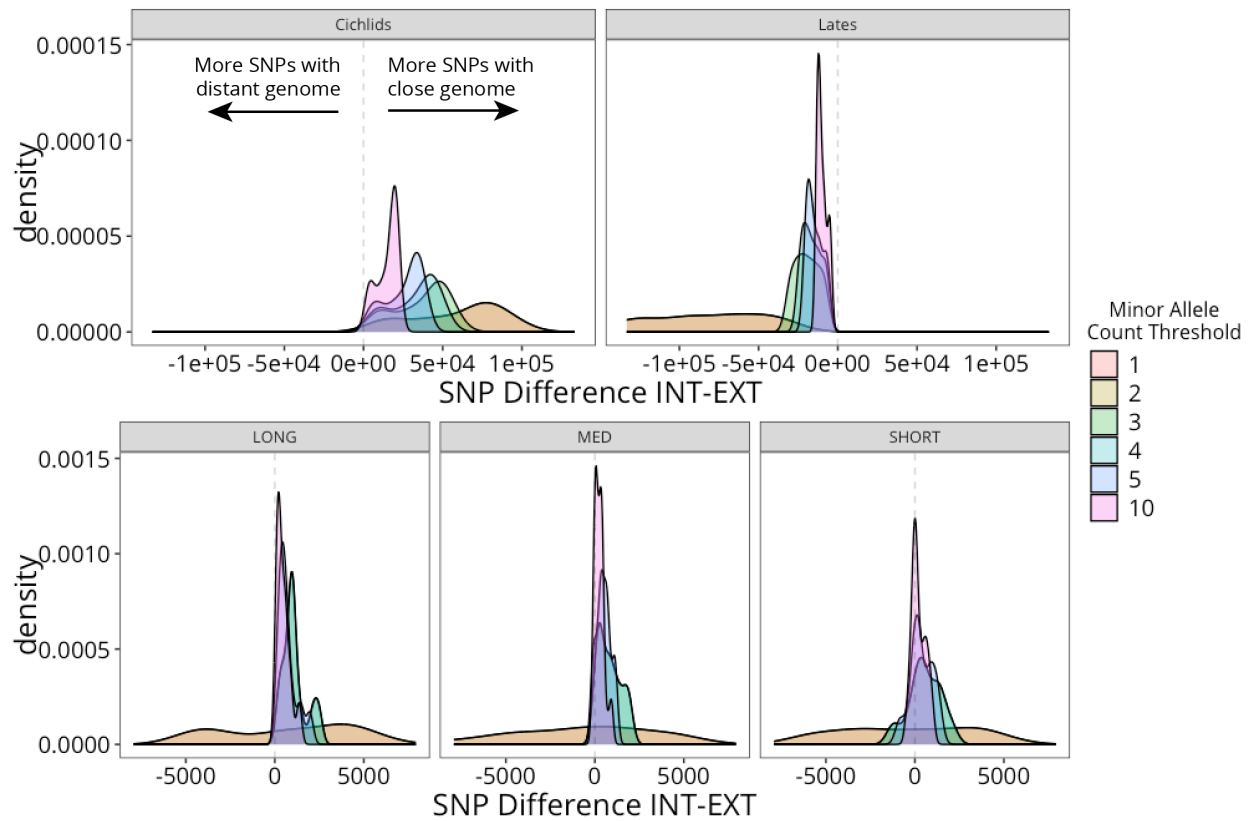


Fig. S29. A comparison of data set size (i.e., number of SNPs) between data sets differing only in the reference genome identity. For each plot, density plots are shown for the difference between number of SNPs with the close reference genome ("INT") and the number of SNPs with the distant reference genome ("EXT"); data sets are colored by the minor allele count threshold used in filtering. For all three ILS levels in the simulations and the trophine clade, data sets aligned to the close reference genome generally had more SNPs, while the opposite was true for the *Lates* data sets. At low MAC thresholds in the simulations, there was less of a clear trend in which data sets had more SNPs.

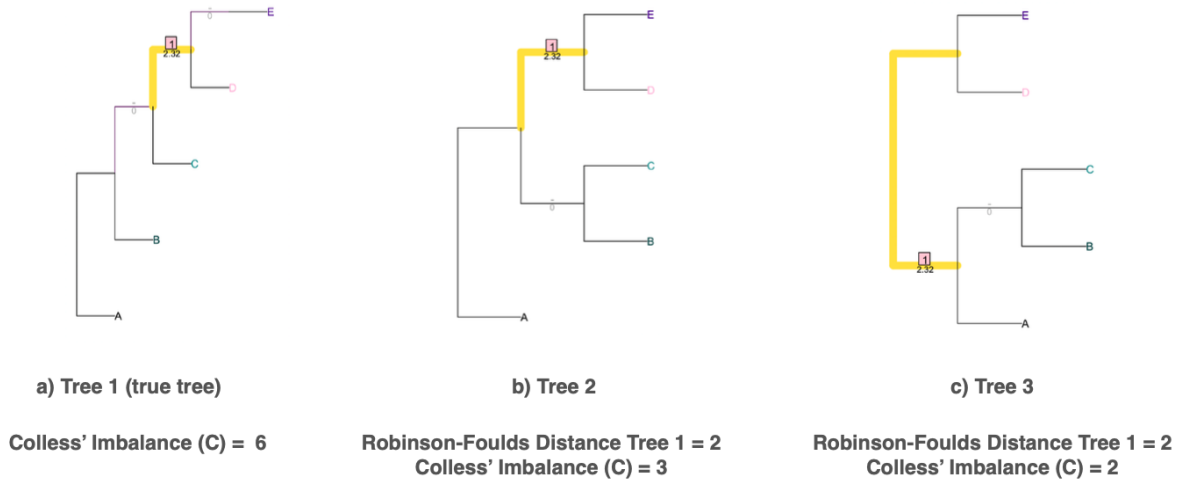


Fig. S30. The imbalance of a topology and its distance with respect to the true topology can vary independently. a) The 'true' topology used to compare Trees 2 (b) and 3 (c) when calculating Robinson-Foulds (RF) distances. Highlighted (yellow) branches indicate shared splits between the focal tree and Tree 1 (i.e., the split isolating clade DE). For Tree 2, the split CB is unique compared to Tree 1 while the split CDE in Tree 1 is unique compared to Tree 2, resulting in RF distance = $1 + 1 = 2$. Similarly, for Tree 3 the split CB is unique compared to Tree 1, which again has the unique split CDE resulting in RF distance = $1 + 1 = 2$. Note that since RF distances are computed excluding the root edge the most basal splits are excluded. The pectinate topology of Tree 1 has the highest Colless imbalance score (C) of $(4 - 1) + (3 - 1) + (2 - 1) = 6$. Tree 2 has a $C = (4 - 1) = 3$ and Tree 3 has a $C = (3 - 2) + (2 - 1) = 2$.

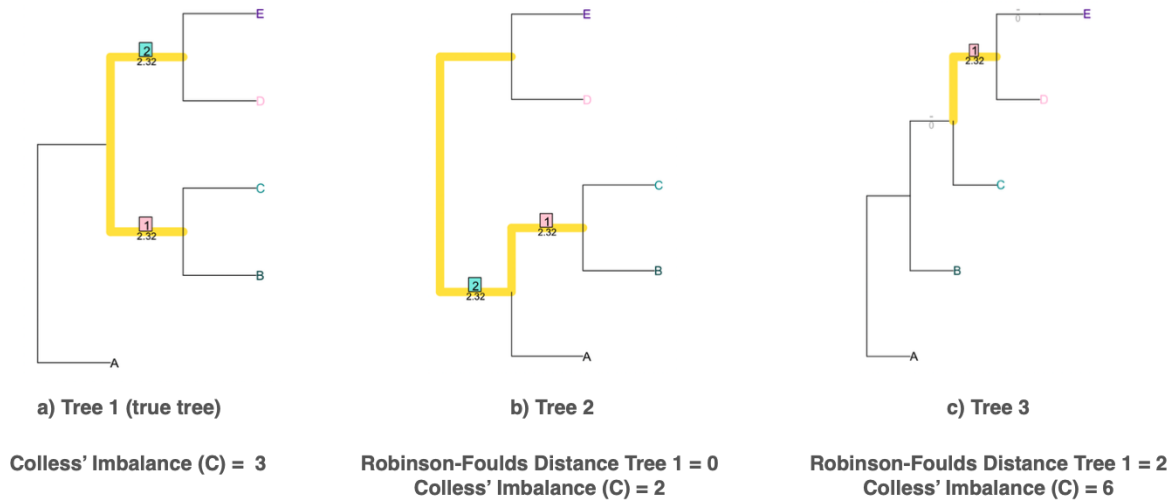


Fig. S31. The imbalance of a topology and its distance with respect to the true topology can vary independently. As before (Fig. S30), highlighted (yellow) branches indicate shared splits between the focal tree and Tree 1 (i.e., the split isolating clade DE). For Tree 2, no splits are unique compared to Tree 1 (both CB and DE are shared and ABC in Tree 2 involves the root edge, and is thus ignored), resulting in RF distance = $0 + 0 = 0$. For Tree 3 the split CDE is unique compared to Tree 1, which has the unique split CB resulting in RF distance = $1 + 1 = 2$. The pectinate topology of Tree 3 has the highest Colless imbalance score (C) of $(4 - 1) + (3 - 1) + (2 - 1) = 6$. Tree 1 has a $C = (4 - 1) = 3$ and Tree 2 has a $C = (3 - 2) + (2 - 1) = 2$.

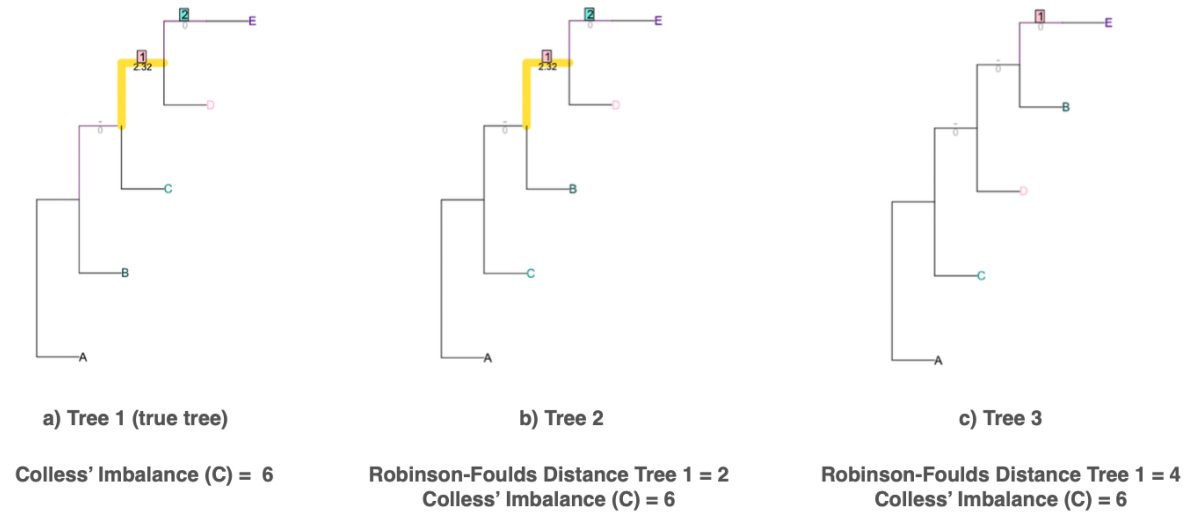


Fig. S32. The imbalance of a topology and its distance with respect to the true topology can vary independently. As before (Fig. S30), highlighted (yellow) branches indicate shared splits between the focal tree and Tree 1 (i.e. the split isolating clade DE for Tree 2 and Tree 1). For Tree 2, the split BDE is unique compared to Tree 1 while the split CDE in Tree 1 is unique compared to Tree 2, resulting in RF distance = $1 + 1 = 2$. For Tree 3, no splits are shared (minus the root edge split, which again is ignored) with Tree 1 resulting in RF distance = $2 + 2 = 4$. All three topologies are pectinate and have the highest Colless imbalance score (C) of $(4 - 1) + (3 - 1) + (2 - 1) = 6$.