Supplementary figures and tables

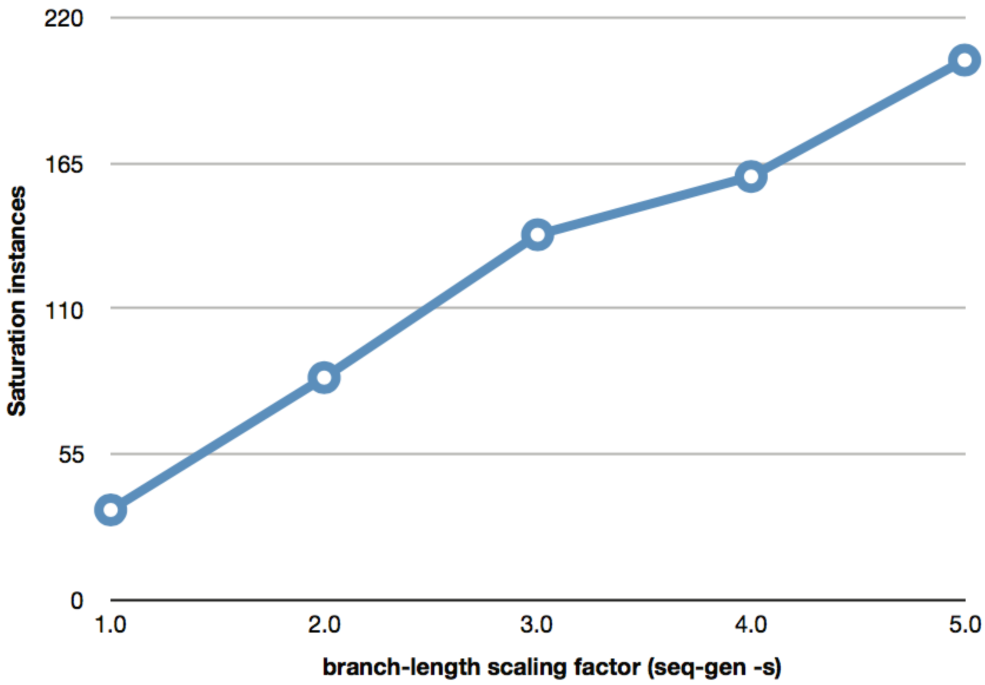


Figure S1. Relationship between the branch length scaling factor and saturation. We ran five instances of Seq-Gen on a six-taxa bifurcating tree, incrementing the branch-length scaling factors. Saturation instances are the number of times an amino acid position changed from one amino acid to another and back. The Y axis is an underestimate since only changes at internal nodes are considered.

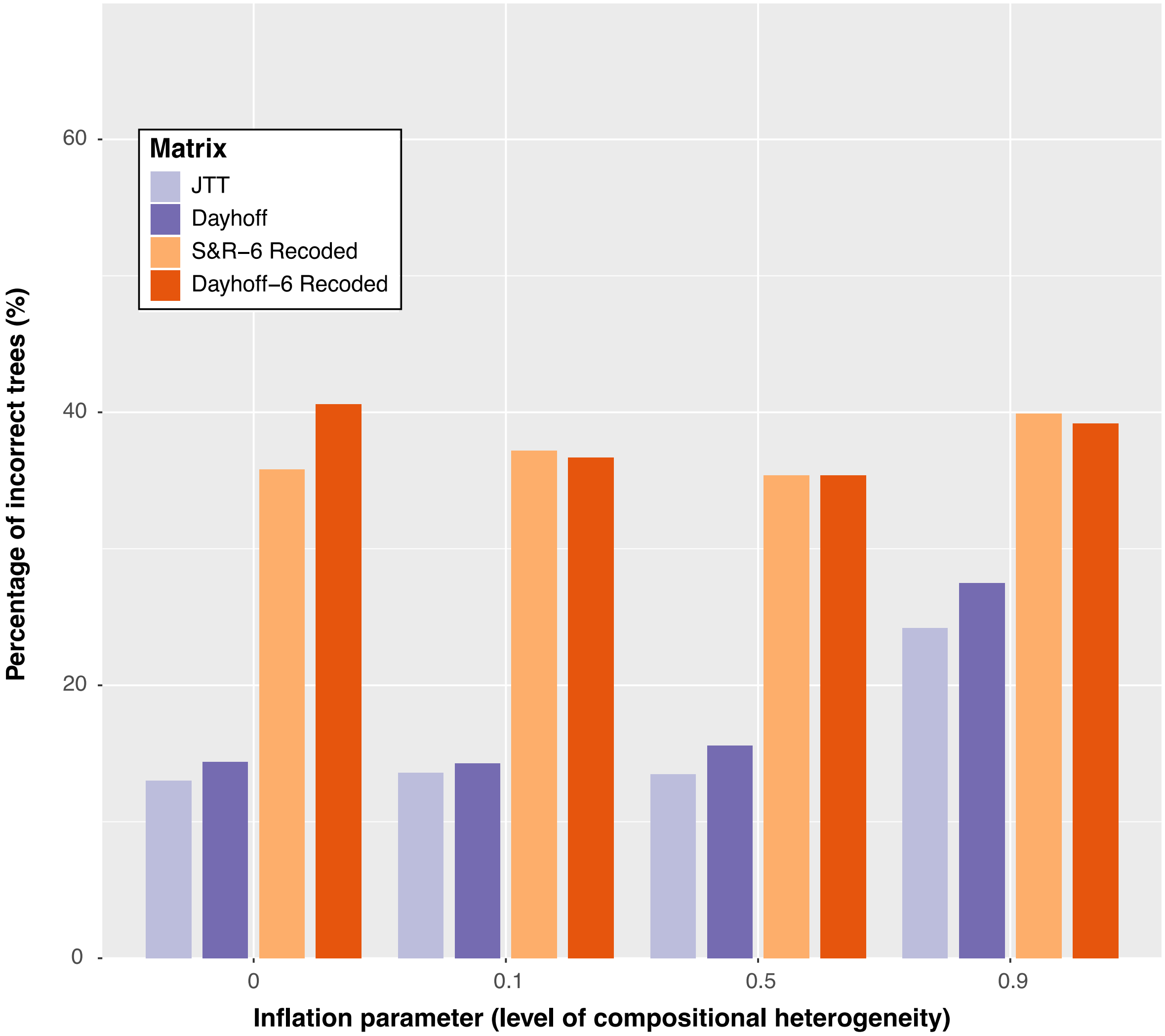


Figure S2. Six-state recoding methods produce more incorrect trees across levels of introduced compositional heterogeneity. Hypothetical tree 0.002 was used for data simulations (Fig. 1a). The inflation parameter set to zero indicates no compositional heterogeneity was introduced into the dataset. Percentage of incorrect trees is out of 1000 trees. Incorrect trees were those that did not reconstruct a monophyletic group of taxa from clade-A and clade-B and monophyletic group of taxa from clade-C and clade-D.

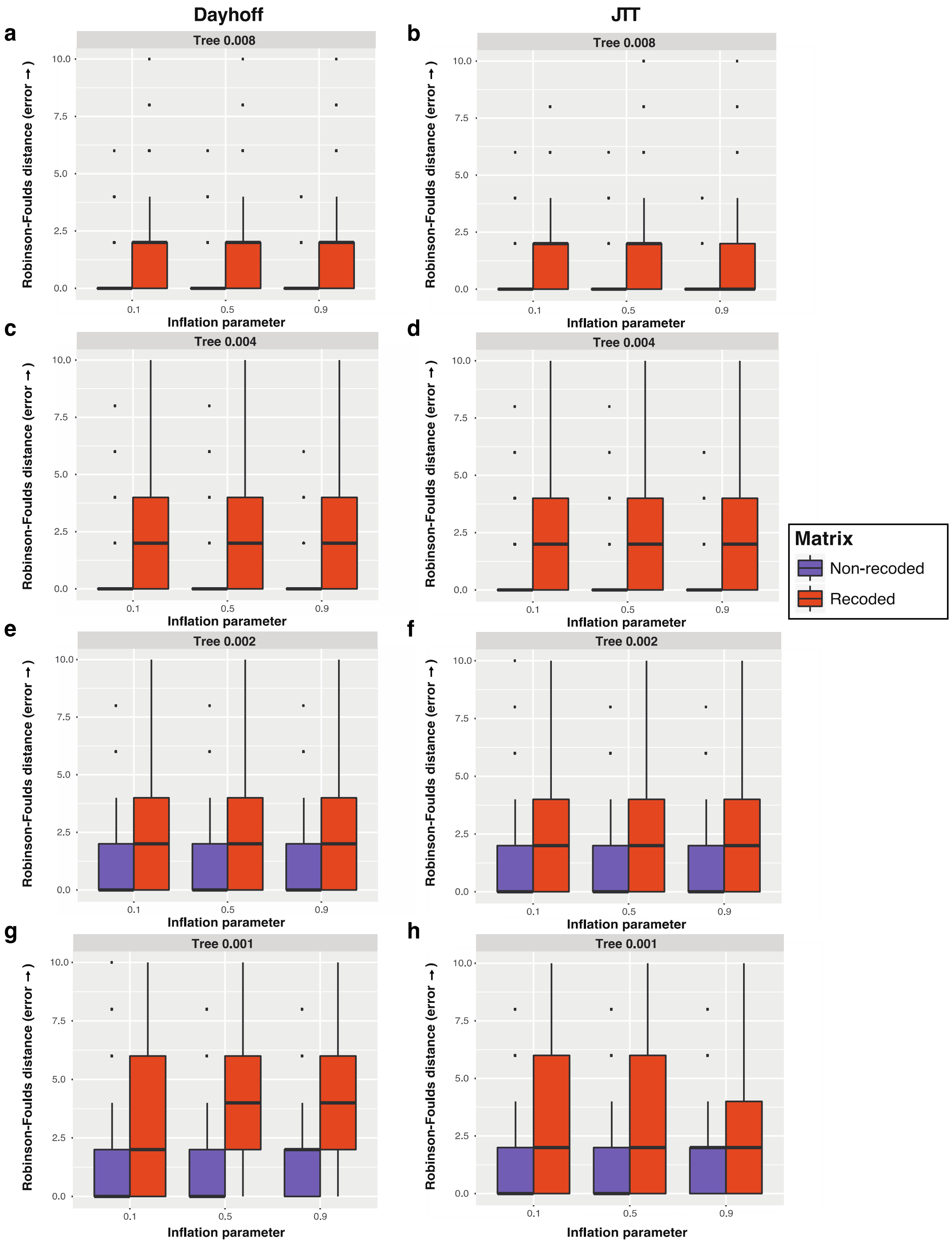


Figure S3. Six-state recoding methods produce more incorrect trees under increasing levels of compositional heterogeneity (inflation parameter settings). Trees were reconstructed either using Dayhoff and Dayhoff 6-state recoding (a, c, e, g) or JTT and S&R 6-state recoding (b, d, f, h). Robinson-Foulds distances were calculated for 1,000 runs for each inflation parameter setting. **(a-b)** Datasets were simulated over hypothetical tree 0.008. **(c-d)** Datasets were simulated over hypothetical tree 0.004. **(e-f)** Datasets were simulated over hypothetical tree 0.002. **(g-h)** Datasets were simulated over hypothetical tree 0.001.

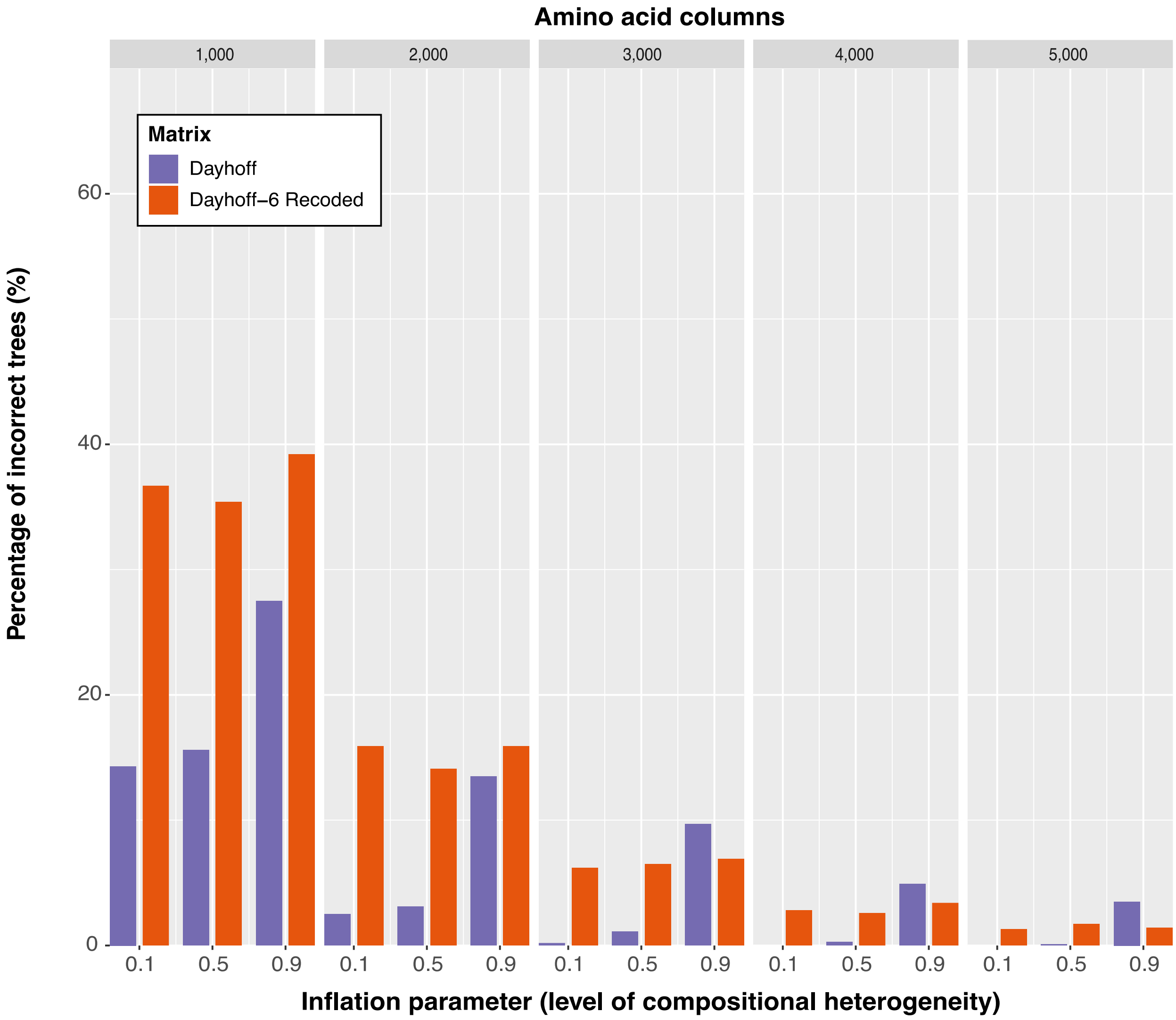


Figure S4. Recoding produces fewer incorrect trees than non-recoding only under the highest level of compositional heterogeneity in larger datasets. Sequences were simulated on hypothetical tree 0.002 (Fig.1a). Incorrect trees were those that did not reconstruct a monophyletic group of taxa from clade-A and clade-B and monophyletic group of taxa from clade-C and clade-D; percentage of incorrect trees is out of 1000 trees.

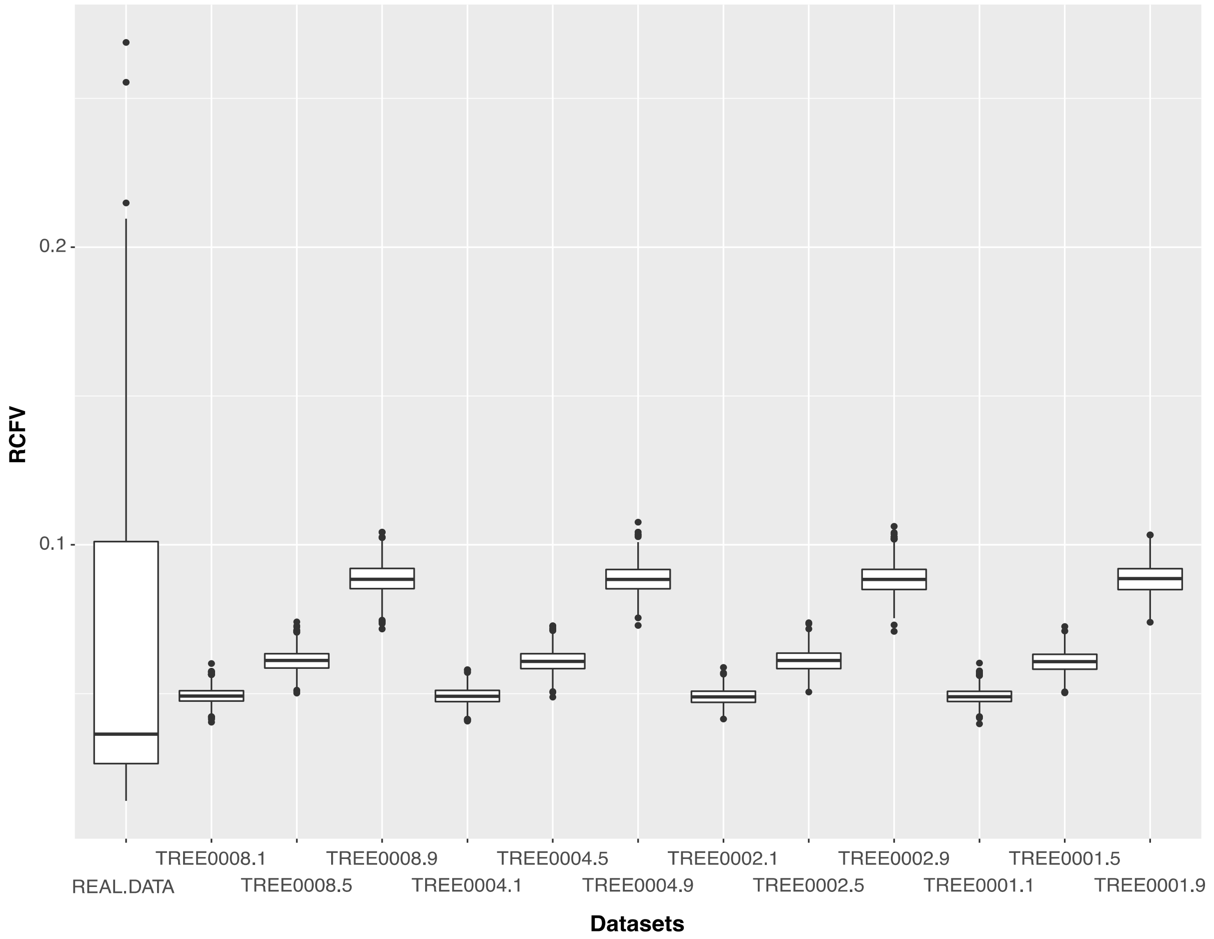


Figure S5. Most of the readily available datasets from Table 1 have lower relative compositional frequency variability (RCFV) scores compared to datasets simulated under the highest level of compositional heterogeneity (inflation parameter set to 0.9). The value following the tree name is the inflation parameter setting (e.g., .1 indicates the inflation parameter was set to 0.1 and data was simulated on hypothetical tree 0.008 in TREE0008.1). Real data used in the analysis are listed in Table S4. Outliers in REAL.DATA result from analyses on single gene datasets.

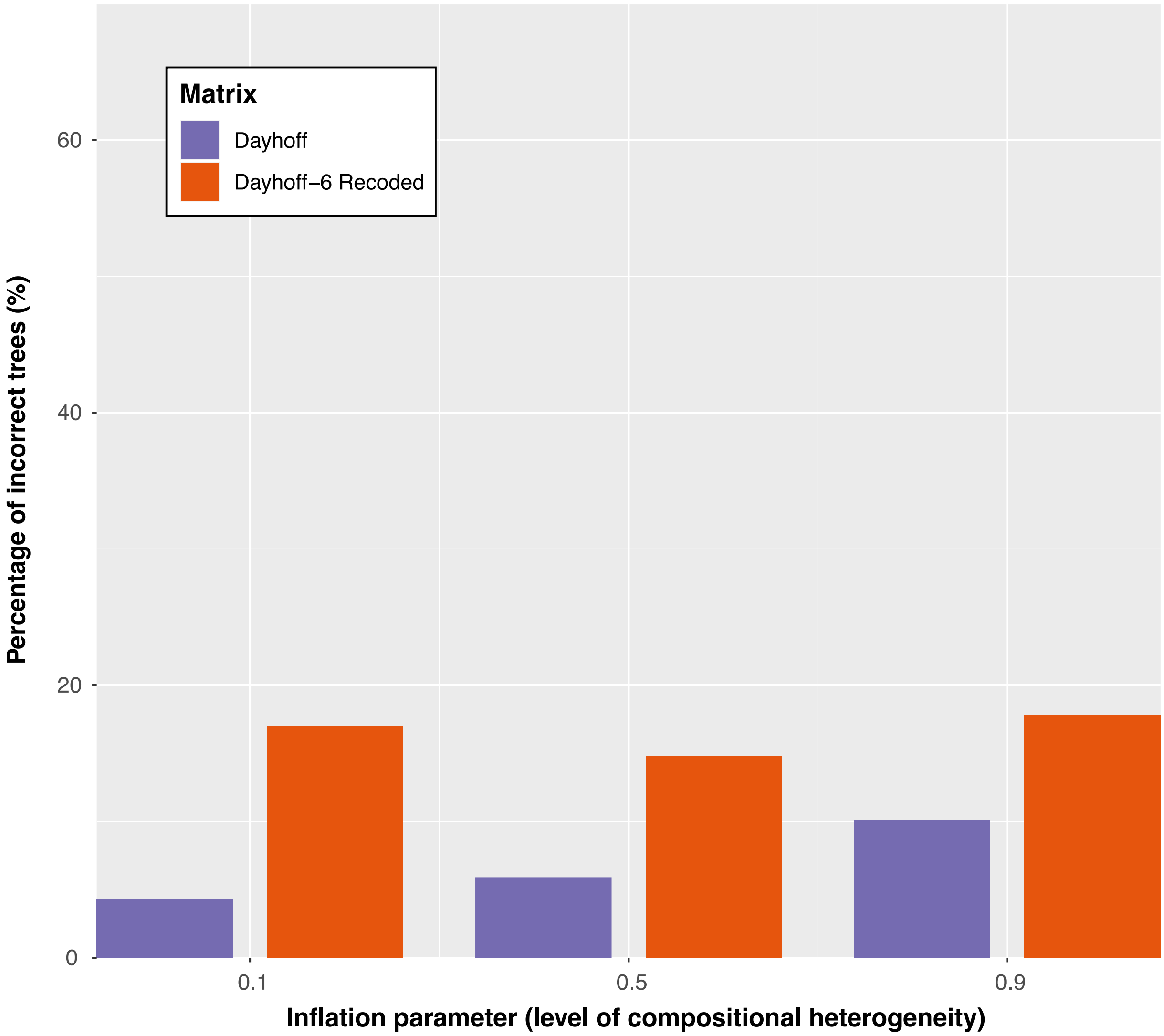


Figure S6. Six-state recoding yields more incorrect trees than non-recoding in datasets produced from an unbalanced tree. Data was simulated over the Chang topology. Incorrect trees were those that did not reconstruct a monophyletic Metazoa and clade including Porifera and ParaHoxozoa (i.e., Placozoa, Cnidaria, and Bilateria). The percentage of incorrect trees is out of 1000 trees.

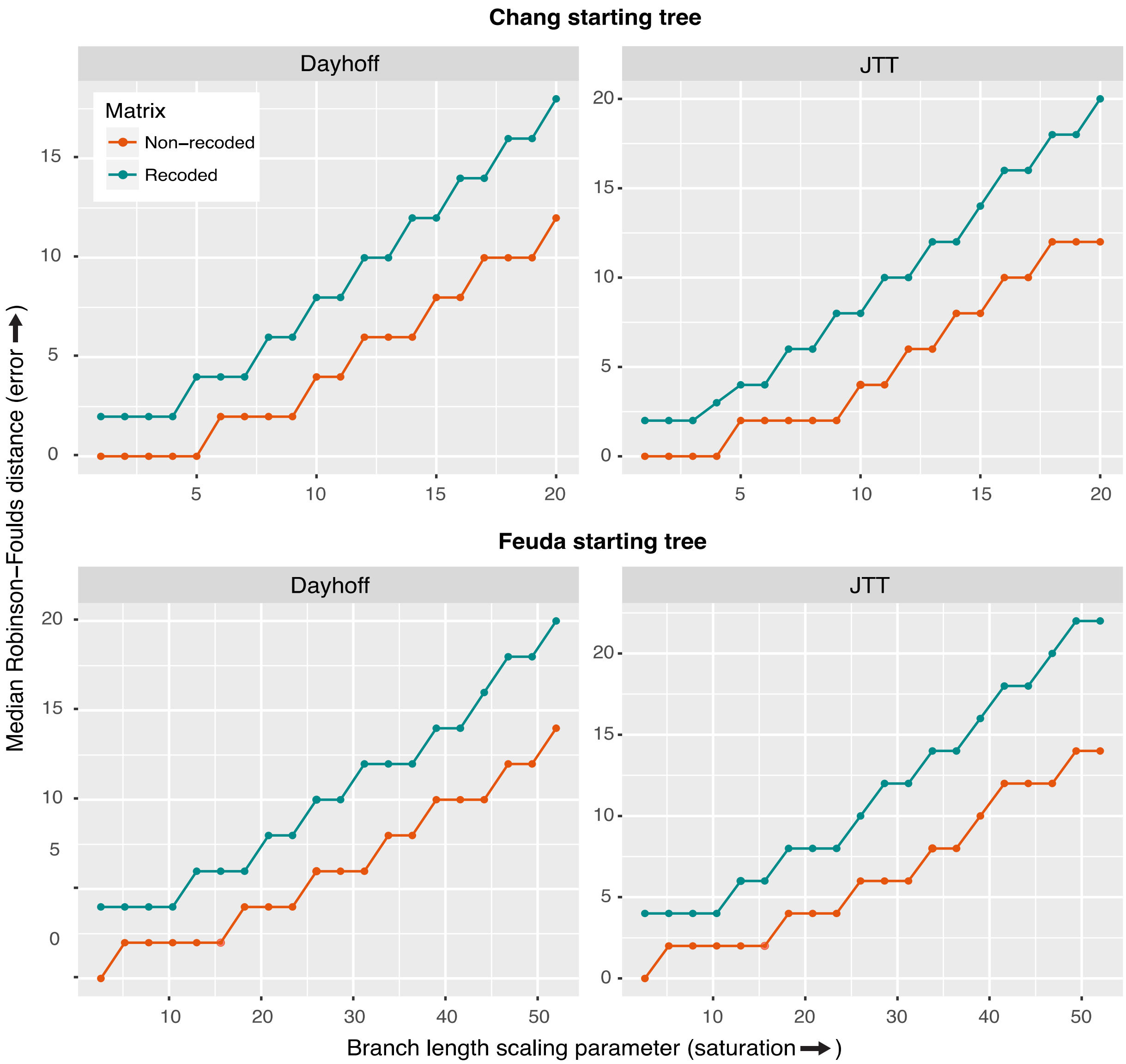


Figure S7. Median Robinson-Foulds distances for non-recoded and recoded datasets across a gradient of saturation levels. Datasets simulated over: **(a)** Chang topology using the Dayhoff model. **(b)** Chang topology under the JTT model. **(c)** Feuda topology under the Dayhoff model. **(d)** Feuda topology using the JTT model.

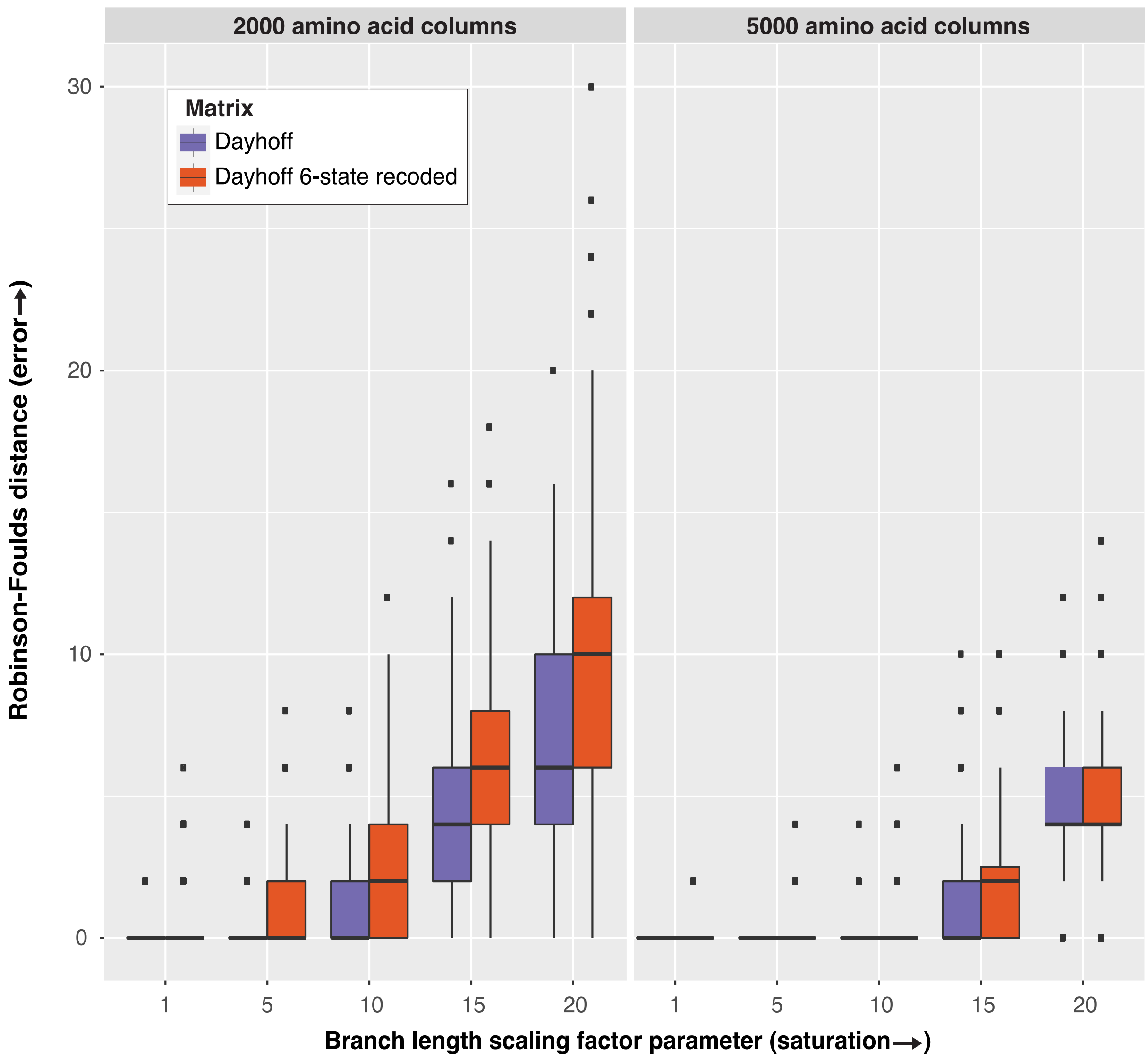


Figure S8. Recoding produces more errors under increasing levels of saturation despite increases in data size. Additionally, these results show that the effects of saturation become greatly reduced with larger datasets. Data was simulated using 2,000 and 5,000 amino acid columns on the Chang topology under the Dayhoff model. Robinson-Foulds distances were calculated for 1,000 runs for each branch length scaling factor parameter.

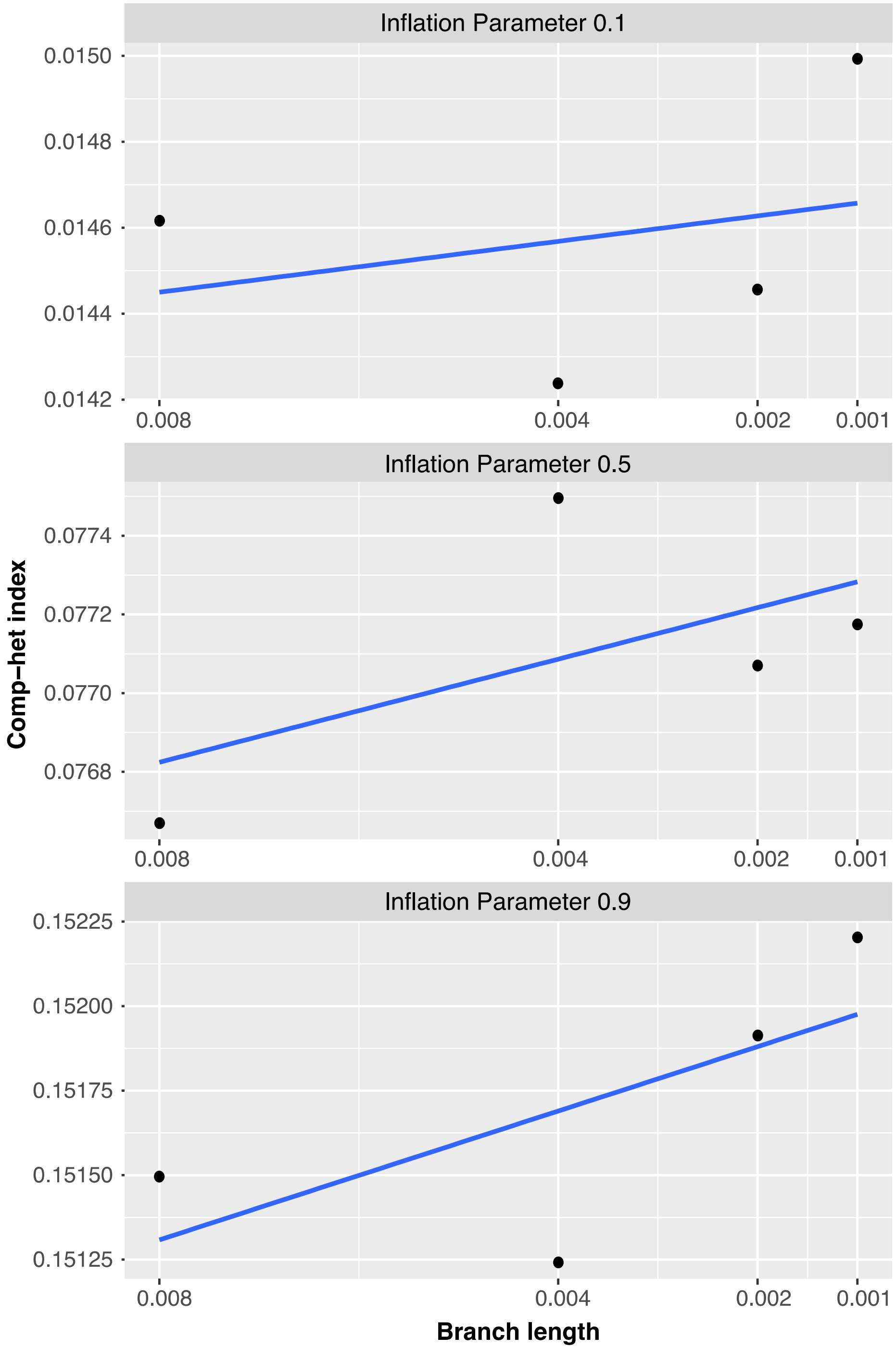


Figure S10. Length of the stem branches of the AB and CD clades (highlighted in orange in Figure 1a) are weakly correlated with comp-het indices. Comp-het indices were calculated by subtracting the mean absolute difference in amino acid frequencies of taxa with homogeneous composition from taxa with heterogeneous composition across 1,000 replicates (see details in supplementary analyses). A limitation of this analysis was the small sample size of data points.

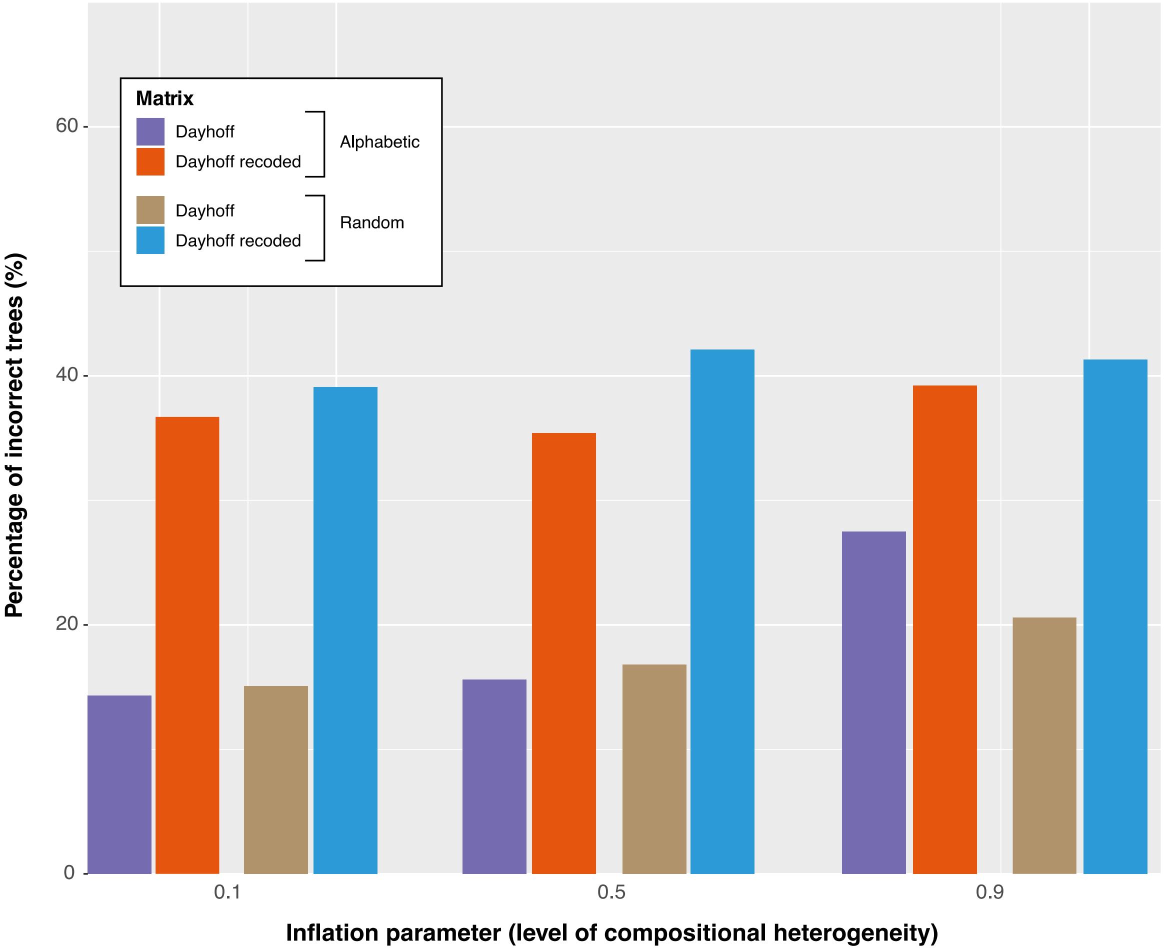


Figure S11. Differences between our pairing strategy and the random pairing strategy to simulate compositional heterogeneity. Sequences were simulated on hypothetical tree 0.002 (Fig.1a). Incorrect trees were those that did not reconstruct a monophyletic group of taxa from clade-A and clade-B and a monophyletic group of taxa from clade-C and clade-D; percentage of incorrect trees is out of 1000 trees. For each of the inflation parameters, our pairing strategy led to results that were more favorable to recoding than to non-recoding.

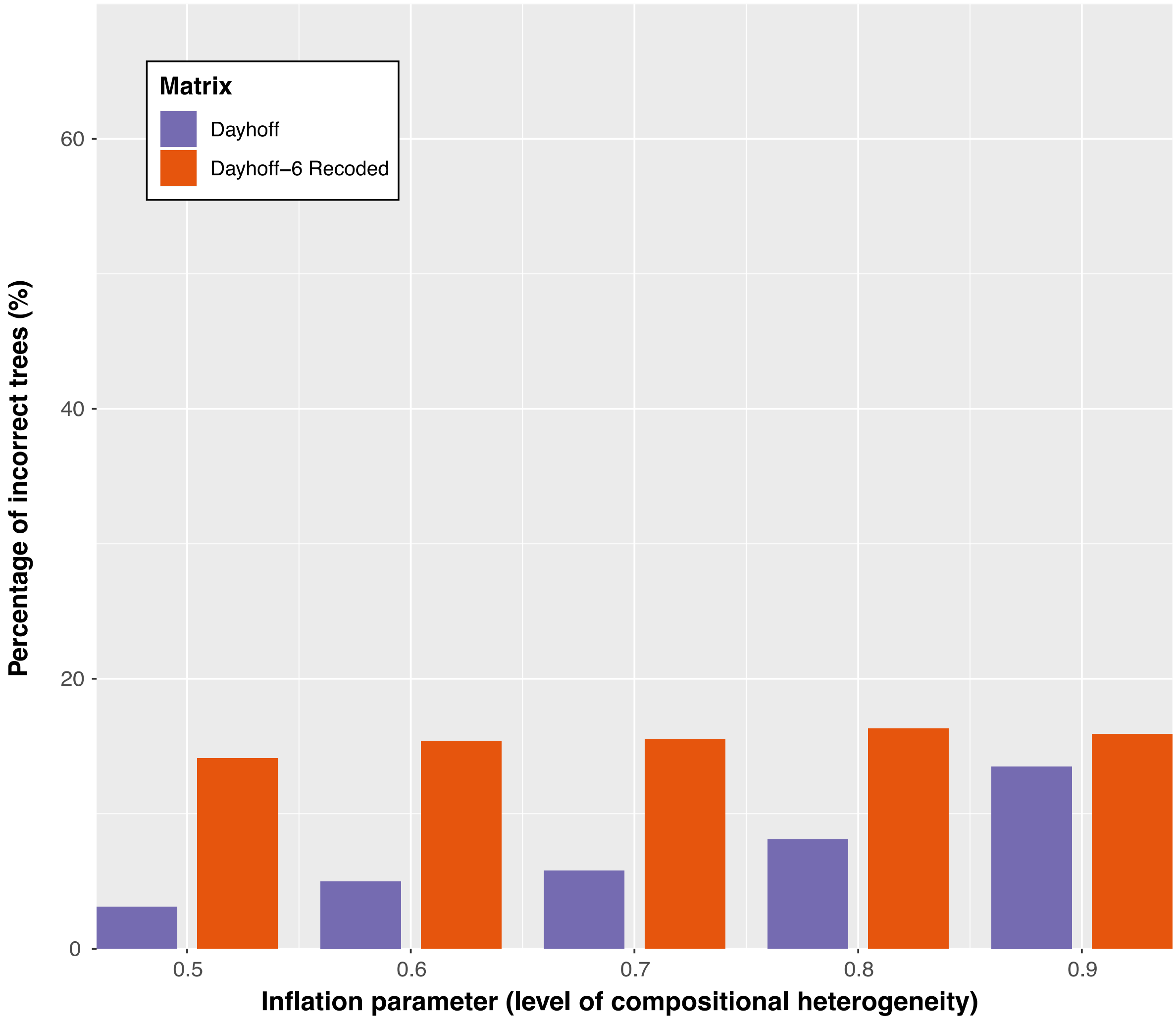
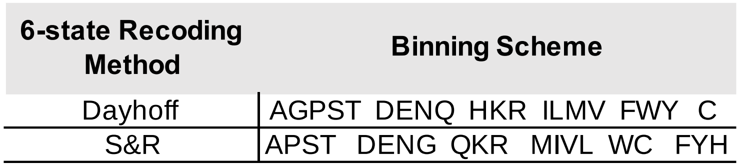


Figure S12. Under non-recoding, the number of incorrect trees gradually increases with increasing compositionally heterogeneity. Sequences were simulated on hypothetical tree 0.002 (Fig.1a) and were made up of 2,000 columns. Incorrect trees were those that did not reconstruct a monophyletic group of taxa from clade-A and clade-B and monophyletic group of taxa from clade-C and clade-D; percentage of incorrect trees is out of 1000 trees.

Inflation Paramter	Clades	Frequency
0.1	clade-A,clade-C	0.081, 0.069, 0.035, 0.051, 0.019, 0.04, 0.058, 0.066, 0.024, 0.075, 0.083, 0.074, 0.03, 0.041, 0.038, 0.051, 0.058, 0.008, 0.028, 0.071
0.5	clade-A,clade-C	0.111, 0.095, 0.025, 0.037, 0.026, 0.054, 0.037, 0.063 ,0.033, 0.102, 0.053, 0.048, 0.04, 0.055, 0.031, 0.037, 0.079, 0.011, 0.019, 0.044
0.9	clade-A,clade-C	0.141, 0.12, 0.014, 0.022, 0.032, 0.068, 0.015, 0.061, 0.042, 0.129, 0.023, 0.023, 0.051, 0.07, 0.025, 0.023, 0.101, 0.013, 0.01, 0.017
0.1, 0.5, 0.9	clade-B, clade-D	0.074, 0.063, 0.038, 0.055, 0.017, 0.036, 0.063, 0.067, 0.022, 0.068, 0.09, 0.08, 0.027, 0.037, 0.04, 0.055, 0.053, 0.007, 0.03, 0.078

Table S1. Amino acid frequencies applied to particular clades under each inflation parameter in compositional heterogeneity simulations. Frequencies are applied in the following order to amino acids and were chosen based on standard input for phylogenetic programs: ARNDCQEGHILKMFPSTWYV.



6-state Recoding Method	Binning Scheme
Dayhoff	AGPST DENQ HKR ILMV FWY C
S&R	APST DENG QKR MIVL WC FYH

Table S2. Binning schemes for Dayhoff and S&R 6-state recoding.

Dayhoff recoding	Binning scheme	Score
9-state	**DEHNQ ILMV FY AST KR G P C W**	**47**
	EHNQ IMTV FY ASL KR G P C W	23
	ILMV DEQ AGPST KN FY H C W R	37
	ILMVT DENQ APS FY KR H G C W	37
12-state	**DEQ MLIV FY G A P S T N KHR W C**	**35**
	DEQ M LIV FY G APST N K H R W C	27
	DEQ M LIV FY GAPS T N K H R W C	26
	DEQ M LIV FY GAPT S N K H R W C	24
	DEQ MLIV FY GAP T S N K H R W C	31
	D E Q MLIV FY GAP S T N KHR W C	29
15-state	**DEQ ML IV FY G A P S T N K H R W C**	**22**
	ML IV G A P S T DE Q N K H RFY W C	10
	ML IV G A P S T DE Q N K H R FYW C	18
	ML IV G A P S T DEQ N K H R FY W C	22
	ML IV G A P S T DEN Q K H R FY W C	21
	ML I V GAP S T DE Q N K H R FY W C	15
	ML IVG A P S T DE Q N K H R FY W C	14
	ML IV G A P S T DEN Q K H R FY W C	21
	ML IV G A P S T D E Q N KHR FY W C	20
	ML I V G A P S T DE Q N KHR FY W C	19
18-state	IV ML G A P S T D E Q N H K R F Y W C	8
	IV M L G A P S T D E Q N H K R F Y WC	-4
	IV M L G A P S T DE Q N H K R F Y W C	7
	IV M L G A P S T D E QN H K R F Y W C	5
	IV ML G A P S T D E Q N H K R F Y W C	8
	MV IL G A P S T D E Q N H K R F Y W C	4
	MI VL G A P S T D E Q N H K R F Y W C	4
	ML IV G A P S T D E Q N H K R F Y W C	8
	ML I V GA P S T D E Q N H K R F Y W C	5
	ML I V G AP S T D E Q N H K R F Y W C	5
	ML I V G A P ST D E Q N H K R F Y W C	5
	ML I V G A P S T DE Q N H K R F Y W C	7
	ML I V G A P S T D E QN H K R F Y W C	5
	ML I V G A P S T D EQ N H K R F Y W C	6
	ML I V G A P S T D E Q N H K RF Y W C	0
	ML I V G A P S T D E Q N H K R FY W C	**11**
	ML I V G A P S T D E Q N HK R F Y W C	4
	ML I V G A P S T D E Q N H KR F Y W C	7
	ML I V G A P S T D E Q N K HR F Y W C	6

Table S3. Binning schemes tested for optimization of substitution scores based on the Dayhoff (PAM 250) log odds matrix. The best scoring schemes are bolded.

Tree	Inflation parameter	Erroneous non-recoded trees	Erroneous recoded trees	P-value
0.008	0.1	1	20	3.38E-05
	0.5	1	25	2.52E-06
	0.9	2	30	7.43E-07
0.004	0.1	21	164	< 2.2e-16
	0.5	21	169	< 2.2e-16
	0.9	65	180	2.03E-13
0.002	0.1	143	367	< 2.2e-16
	0.5	156	354	< 2.2e-16
	0.9	275	392	5.89E-06
0.001	0.1	336	551	5.24E-13
	0.5	393	581	1.70E-09
	0.9	518	590	0.03054

Table S4. P-values of chi-squared tests between incorrect trees solved with Dayhoff and Dayhoff 6-state recoding in compositional heterogeneity analyses (α = 0.002).

Tree	Inflation parameter	Erroneous non-recoded trees	Erroneous recoded trees	P-value
0.008	0.1	1	18	9.62E-05
	0.5	0	30	4.32E-08
	0.9	3	34	3.46E-07
0.004	0.1	21	144	< 2.2e-16
	0.5	26	157	< 2.2e-16
	0.9	62	163	1.66E-11
0.002	0.1	136	372	< 2.2e-16
	0.5	135	354	< 2.2e-16
	0.9	242	399	5.61E-10
0.001	0.1	322	551	9.16E-15
	0.5	367	556	4.94E-10
	0.9	484	562	0.01588

Table S5. P-values from chi-squared tests between incorrect trees solved with JTT and S&R 6-state recoding in compositional heterogeneity analyses ($\alpha = 0.002$).

Data size (amino acid columns)	Branch length scaling factor parameter	P-value
2000	1	2.20E-16
	5	2.20E-16
	10	2.20E-16
	15	2.20E-16
	20	2.20E-16
5000	1	2.27E-02
	5	2.20E-16
	10	0.0007505
	15	2.20E-16
	20	1.44E-03

Table S6. P-values from t-tests of Robinson-Foulds distances between recoded and non-recoded datasets in saturation analyses (α = 0.005).

Citation	Dataset	RCFV
Williams et al. (2011)	Williams2011_pcna_aln.fa	0.26879
Williams et al. (2011)	Williams2011_tf2b_aln.fa	0.25542
Andersson et al. (2006)	Andersson2006_nagB135_229.phy	0.21487
Williams et al. (2011)	Williams2011_fen_aln.fa	0.20957
Andersson et al. (2006)	Andersson2006_12862_2005_212_MOESM8_ESM.ande.fa	0.20533
Andersson et al. (2006)	Andersson2006_12862_2005_212_MOESM7_ESM.ande.fa	0.19984
Domman et al. (2015)	Domman2015_GlgA_Ball_2013.fa	0.18584
Williams et al. (2011)	Williams2011_rnap2_aln.fa	0.16266
Andersson et al. (2006)	Andersson2006_12862_2005_212_MOESM6_ESM.ande.fa	0.14764
Moore et al. (2019)	Moore2019_Figure3.fa	0.12027
Moore et al. (2019)	Moore2019_Figure2.fa	0.11333
Moore et al. (2019)	Moore2019_Figure1.fa	0.11254
Lasek-Nesselquist & Gogarten (2013)	Lasek-Nesselquist2013_aeb1.fa	0.10456
Andersson et al. (2006)	Andersson2006_nagB13_560.phy	0.10287
Andersson et al. (2006)	Andersson2006_12862_2005_212_MOESM9_ESM.ande.fa	0.10047
Song et al. (2016)	Song2016_data_aa.fa	0.09627
Puttick et al. (2018)	Puttick2018_aa_566270.phy	0.05744
Manzano-Marín et al. (2018)	Manzano-Marin2018_file_S2.phylip	0.05339
Feuda et al. (2017)	Feuda2017_WhelanD20_AA.phy	0.05267
Laumer et al. (2018)	Laumer2018_Tplx_phylo_p4failed_d0.5.fa	0.05242
Simion et al. (2017)	Simion2017_supermatrix_90sp_102464pos_heterop70.phy	0.04963
Simion et al. (2017)	Simion2017_supermatrix_90sp_136618pos_heterop60.phy	0.04741
Leliaert et al. (2016)	Leliaert2016_1_1459075542_infile.fa	0.04467
Feuda et al. (2017)	Feuda2017_Whelan_D16_Choanoflagellata_outlier_excluded.phy	0.04466
Laumer et al. (2018)	Laumer2018_Tplx_BUSCOeuk.fa	0.04364
Laumer et al. (2018)	Laumer2018_Tplx_phylo_d1.fa	0.04193
Laumer et al. (2018)	Laumer2018_Tplx_phylo_p4pass_d0.5.fa	0.04110
Eitel et al. (2018)	Eitel2018_dataset_2.phy	0.03659
Lemieux et al. (2014)	Lemieux2014_Amino_acid_dataset.fa	0.03616
Borowiec et al. (2015)	Borowiec2015_Best108.phy	0.03597
Laumer et al. (2019)	Laumer2019_spiralia_BMGE.phylip	0.03500
Lemer et al. (2019)	Lemer2019_Matrix1.fa	0.03358
Lemer et al. (2019)	Lemer2019_Matrix3i.fa	0.03307
Ballesteros et al. (2019)	Ballasteros2019_Matrix_5.fa	0.03292
Schwentner et al. (2018)	Schwentner2018_Matrix5_aa.fa	0.03119
Ballesteros et al. (2019)	Ballasteros2019_Matrix_2.fa	0.03070
Feuda et al. (2017)	Feuda2017_Chang_AA.phy	0.02997
Ballesteros et al. (2019)	Ballasteros2019_Matrix_1.fa	0.02994
Schwentner et al. (2018)	Schwentner2018_Matrix1_aa.fa	0.02830
Marlétaz et al. (2019)	Marletaz2019_Concat-Tc111217-proto.phy	0.02812
Schwentner et al. (2018)	Schwentner2018_Matrix4_aa.fa	0.02651
Laumer et al. (2019)	Laumer2019_mtz_UPhO_pBMGE.phylip	0.02651
Philippe et al. (2019)	Philippe2019_all_genes_ordered_by_monophyly.fa	0.02651
Rota-Stabelli et al. (2013)	Rota-Stabelli2013_aa.fa	0.02640
Marlétaz et al. (2019)	Marletaz2019_Concat-Tc111217-broad.phy	0.02631
Schwentner et al. (2018)	Schwentner2018_Matrix6_aa.fa	0.02549
Schwentner et al. (2018)	Schwentner2018_Matrix3_aa.fa	0.02522
Schwentner et al. (2018)	Schwentner2018_Matrix2_aa.fa	0.02447
Laumer et al. (2019)	Laumer2019_nonbilateria_MARE_BMGE.phy	0.02444
Marlétaz et al. (2019)	Marletaz2019_Concat-Tc111217-red.phy	0.02336
Marlétaz et al. (2019)	Marletaz2019_Concat-Tc111217-strin.phy	0.02336
Lozano-Fernandez et al. (2019)	Lozano-Fernandez_2019_MatrixA.phy	0.02305
Marlétaz et al. (2019)	Marletaz2019_Concat-Tc111217-cnid.phy	0.02286
Laumer et al. (2019)	Laumer2019_nonbilateria_MARE_cho_BMGE.phy	0.01997
Wolfe et al. (2019)	Wolfe2019_1_aa_410.fa	0.01481
Schwentner et al. (2018)	Schwentner2018_Matrix7_aa.fa	0.01400

Table S7. Datasets and RCFV scores from 25 publications that use 6-state recoding.

	Real Data
TREE0008.1	0.067
TREE0008.5	8.30E-06
TREE0008.9	4.10E-06
TREE0004.1	0.08
TREE0004.5	8.90E-06
TREE0004.9	4.30E-06
TREE0002.1	0.087
TREE0002.5	8.70E-06
TREE0002.9	4.20E-06
TREE0001.1	0.076
TREE0001.5	8.80E-06
TREE0001.9	4.10E-06

Table S8. Pairwise Wilcoxon rank sum test p-values between RCFV scores from real data compared to scores from simulated data. Tree names indicate comparisons of RCFV scores from data simulated over each tree in Figure 1a and inflation parameter applied. The value following the tree name is the inflation parameter setting (e.g., .1 indicates the inflation parameter was set to 0.1 and data was simulated on hypothetical tree 0.008 in TREE0008.1). Real data used in the analysis are listed in Table S7.