Supplementary Analyses

1. Datasets with no Introduced Compositional Heterogeneity

For each hypothetical tree used for simulation (Fig. 1a), we simulated 1,000,000 datasets with no introduced compositional heterogeneity (inflation parameter set to 0) to generate a null distribution of datasets to which we compare data with introduced compositional heterogeneity (i.e., simulated with the inflation parameters set to 0.1, 0.5, and 0.9). To verify that these data displayed compositional heterogeneity between clade-A and clade-C compared to clade-B and clade-D, we computed compositional heterogeneity (comp-het) indices. Comp-het indices were produced by calculating amino acid frequencies from the simulated datasets and summing the absolute difference in frequencies across all replicates. We subtracted the mean absolute difference in amino acid frequencies between clade-A and clade-C and between clade-B and clade-D (clade pairs with homogeneous composition) from the mean absolute difference in amino acid frequencies between clade-A and clade-B, clade-A and clade-D, clade-B and clade-C, as well as clade-C and clade-D (clade pairs with heterogeneous composition).

Datasets with comp-het indices closer to 0 are characterized by low compositional heterogeneity, while datasets with higher comp-het indices are characterized by high compositional heterogeneity (Fig. S10). We counted the number of comp-het indices from the null distribution that were greater than comp-het indices from compositionally heterogeneous datasets to determine significant difference between data composition. There were no cases in which comp-het indices from the null distribution were greater than compositionally heterogeneous datasets (p-value = 0).

We reconstructed maximum-likelihood trees in RAxML for the first 1,000 out of 1,000,000 datasets simulated with no introduced compositional heterogeneity over hypothetical tree 0.002. We applied Dayhoff and JTT models for phylogenetic reconstruction of non-recoded data and the GTR multi-state model for Dayhoff 6-state and S&R 6-state recoded data. Phylogenetic analyses of recoded datasets performed worse than analyses of non-recoded datasets, regardless of the amount of compositional heterogeneity introduced (Figure S2).

2. Effect of Amino Acid Frequencies and Compositional Heterogeneity on Recoding Methods (Random Pairings)

To address whether our choice of amino acid pairing strategy for simulating compositionally heterogeneous data had an effect on the analyses, we ran additional simulations using starting amino acid frequency pairs sampled at random. We simulated 1,000 datasets over hypothetical tree 0.002 with random pairings of amino acid frequencies across replicates for each inflation parameter setting (i.e., 0.1, 0.5, and 0.9) under the GTR model. We performed all simulations using the script comphet.pl in which we implement our compositional heterogeneity algorithm in P4. All simulated sequences were 1,000 amino acids in length. We recoded simulated datasets using Dayhoff 6-state recoding and performed maximum-likelihood analysis in RAxML using the GTR multi-state model for recoded datasets and the Dayhoff model for non-recoded datasets. Our simulations produced a total of 6,000 trees from 3,000 datasets. We scored trees according to the compositional heterogeneity analysis outlined in the main text, where we use the script is_mono.pl to assess whether trees recovered a monophyletic group

including all clade-A and clade-B taxa, and a monophyletic group including all clade-C and clade-D taxa. Our pairing strategy applied in the main text and the random pairing strategy produced similar results in that non-recoding outperformed recoding under all levels of compositional heterogeneity tested (Fig. S11). The biggest differences were the performance of recoding on the 0.5 inflation parameter simulations and the performance of non-recoding on the 0.9 inflation parameter simulations (Fig. S11). In both cases, these differences favored recoding making it unlikely that our paring strategy introduced a bias against recoding.

3. Testing 6-state Recoding Performance on Data Size

To test the effect of data size on Dayhoff 6-state recoding methods, we simulated 12,000 datasets with varying levels of compositional heterogeneity and sequence lengths. We simulated 1,000 datasets on hypothetical tree 0.002 for each combination of four sequence lengths (2,000, 3,000, 4,000, and 5,000 amino acid columns) and three inflation parameters settings (i.e., 0.1, 0.5, and 0.9) under the GTR model. All simulations were performed using the script comphet.pl for implementation of our compositional heterogeneity algorithm in P4. We recoded each simulated dataset with Dayhoff 6-state recoding and reconstructed maximum-likelihood trees in RAxML using the GTR multi-state model for recoded datasets and the Dayhoff model for non-recoded datasets. Our simulations produced a total of 24,000 trees. We scored the trees as outlined in the compositional heterogeneity analyses in the main text, where we use the script is_mono.pl to assess whether trees recovered a monophyletic group that included all clade-A and clade-B taxa, and a monophyletic group that included all clade-C and clade-

D taxa. As sequence length increased, recoding methods outperformed non-recoding methods only under the highest level of compositional heterogeneity (Fig. S4; i.e., inflation parameter = 0.9).

Since there was a substantial increase in the percentage of incorrect trees when the inflation parameter was set to 0.9 in our non-recoded analyses of 2,000-column datasets (Figure S4), we performed additional analyses with 2,000-column data simulated on hypothetical tree 0.002 and the inflation parameters set to 0.6, 0.7, and 0.8 to better understand the dynamic shift. We simulated 1,000 datasets for each inflation parameter tested, totaling to 3,000 datasets. Simulations were performed with the script comphet.pl to introduce compositional heterogeneity in P4 under the GTR model. We recoded each simulated dataset with Dayhoff 6-state recoding and reconstructed maximum-likelihood trees in RAxML using the GTR multi-state model for recoded datasets and the Dayhoff model for non-recoded datasets. Our simulations produced a total of 6,000 trees, which were scored with the script is_mono.pl to assess whether a monophyletic group that included all clade-A and clade-B taxa, and a monophyletic group that included all clade-C and clade-D taxa, were recovered. The number of incorrect trees under non-recoding gradually increased with increasingly compositionally heterogeneous datasets (Fig. S12).

We also simulated 10,000 datasets with increasing levels of saturation and varying sequence lengths. We used Seq-Gen to simulate 1,000 datasets on the Chang topology for a combination of two sequence lengths (i.e., 2,000 and 5,000 amino acid columns) and five branch length scaling factor parameters subset from the saturation analyses in the main text (i.e., 1, 5, 10, 15, 20) under the Dayhoff model. Each dataset was recoded using Dayhoff 6-state recoding and reconstructed in RAxML using the GTR

multi-state model for recoded datasets, as well as the Dayhoff model for non-recoded datasets. We reconstructed a total of 24,000 trees, which we scored using TOPD/FMTS to calculate Robinson-Foulds distances from comparisons to the Chang topology. We used a t-test to determine if there were significant differences in Robinson-Foulds distances between recoded and non-recoded datasets for each combination of dataset size and branch length scaling factor. To correct for multiple t-tests, we applied the Bonferroni correction at which $\alpha = 0.005$. Increases in sequence length minimized the impact of saturation and reduced errors in phylogenetic reconstruction for both recoding and non-recoding methods, but recoding methods performed significantly worse than non-recoding methods in all except the lowest level of saturation in the largest simulated datasets (Fig. S8; Table S4).

4. Comparing Compositional Heterogeneity in Real Data to Simulated Data

Given the decline in performance of larger, non-recoded datasets at very high inflation parameters (Fig. S4), we assessed how simulated levels of compositional heterogeneity compared to naturally occurring compositional heterogeneity present in real datasets. We compiled up to 56 datasets (Table S5) from 25 publications that use 6-state recoding from Table 1. We used the rcfv.pl to calculate the average relative compositional frequency variability (RCFV) score from BaCoCa (Kück & Struck 2014) for each of the simulated datasets in the compositional heterogeneity analysis from the main text and the 56 real datasets in Table S4.

To determine which statistical test would be most appropriate to compare RCFV scores from simulated and real data, we used the Shapiro-Wilk's test in R to test if RCFV

scores were normally distributed with the script normality_test.R. We then applied the Kruskal–Wallis test to determine if there were significant differences between the RCFV scores from simulated and real data. We chose the Kruskal-Wallis test because the RCFV scores were not normally distributed. RCFV scores between real and simulated data significantly differed (p-value < 2.2e-16). To determine which specific simulated datasets significantly differed from the real data, we used the pairwise Wilcoxon rank sum test with the Bonferroni correction for multiple testing. RCFV scores from the real data significantly differed from all simulated datasets except those simulated under the inflation parameter 0.1 (Table S5). We used the script Boxplot_Kruskal_Wallis.R to visualize the data using a box-plot, perform the Kruskal-Wallis test, and the pairwise Wilcoxon rank sum test. All scripts are available in our GitHub repository.

5. **Testing 6-state Recoding Performance on Tree Shape and Compositional Heterogeneity**

To test whether the choice to use a simple, balanced topology in our compositional heterogeneity analyses led to artifactual results, we simulated 1,000 datasets for each inflation parameter setting (i.e., 0.1, 0.5, and 0.9) over the Chang tree using the GTR model. We used the script comphet.pl to implement our compositional heterogeneity algorithm in P4 where we applied one set of amino acid frequencies to the Porifera and Choanoflagellata clades, while applying a different set of amino acid frequencies to the remaining clades on the tree. All simulated sequences were 1,000 amino acids in length. We recoded each dataset using Dayhoff 6-state recoding and reconstructed trees using the GTR multi-state model in RAxML. Phylogenetic trees for

non-recoded datasets were reconstructed using the Dayhoff model in RAxML. We simulated a total of 3,000 datasets and reconstructed 6,000 trees. We scored trees by deep splits as outlined in the main text using the script is_mono.pl to first assess whether trees recovered a monophyletic Metazoa and then to test whether trees recovered a clade including Porifera and ParaHoxozoa (i.e., Placozoa, Cnidaria, and Bilateria). We chose to test these relationships, because the positions of Ctenophora and Porifera are amongst the most contested questions in phylogenetics (Dunn et al. 2008; Hejnol et al., 2009; Philippe et al., 2009; Pick et al., 2010; Ryan et al. 2013; Moroz et al. 2014; Boroweic et al. 2015; Chang et al. 2015; Pisani et al. 2015; Whelan et al. 2015; Telford et al. 2015; Feuda et al. 2017; Shen et al. 2017; Laumer et al. 2018; Laumer et al. 2019; Pett et al. 2019). Under all levels of compositional heterogeneity tested, the results from these analyses were consistent with those derived from symmetric trees (i.e., non-recoding analyses outperformed recoding analyses in all cases; Fig. S6).