Recoding amino acids to a reduced alphabet may increase or decrease phylogenetic accuracy

Supplementary material

May 10, 2022

Authors

Peter G. Foster [1*], Dominik Schrempf [2], Gergely J. Szöllősi [2,3,4], Tom A. Williams [5], Cymon J. Cox [6], T. Martin Embley [7]

[1] Department of Life Sciences, Natural History Museum, London SW7 5BD, UK
[2] Department of Biological Physics, Eötvös Loránd University, 1117 Budapest, Hungary
[3] MTA-ELTE "Lendület" Evolutionary Genomics Research Group, 1117 Budapest, Hungary
[4] Evolutionary Systems Research Group, Centre for Ecological Research, Hungarian Academy of Sciences, 8237 Tihany, Hungary
[5] School of Biological Sciences, University of Bristol, BS8 1TQ, Bristol, UK
[6] Centro de Ciências do Mar, Universidade do Algarve, Gambelas, 8005-319 Faro, Portugal
[7] Biosciences Institute, Centre for Bacterial Cell Biology, Baddiley-Clark Building (room 2.04), Newcastle University, Richardson Road, Newcastle upon Tyne, UK
[*] Corresponding author `p.foster2@gmail.com`

Supplementary material

Contents

Information paucity: Information loss due to recoding matters when the amount of phylogenetic information is limited

Hernandez and Ryan (2021) (hereafter H&R) have made a recent commentary on recoding of amino acid datasets. One part of their study looked at accuracy of recoded datasets in the face of compositional heterogeneity over the tree. H&R focussed on the effect of the decrease in information associated with recoding. The main results for this aspect of their study are presented in their Figure 1b, which shows that recoded analyses are less accurate than unrecoded. They interpret this to mean that while recoded data appears to be more immune to compositional heterogeneity, this is outweighed by the penalty due to the loss of information due to recoding. Here we look more closely at their results for their Tree 0.002 as an example. Results for this tree in H&R Figure 1b are repeated in their Figure S2, with the addition of results with no imparted compositional heterogeneity (inflation parameter 0.0). For simplicity we only look at Dayhoff analysis of simulated protein datasets, and Dayhoff 6-state recoding (H&R additionally used JTT model analyses of simulated protein datasets, and SR-sat, which H&R call S&R-6, recoding).

The tree that they used for their simulations is given on their github site, as

```
our $TREE0002 = '((((A1:0.10,(A2:0.05,A3:0.05):0.05):0.075,(A4:0.10,A5:0.10)
    :0.075):0.008577,((B1:0.10,(B2:0.05,B3:0.05):0.05):0.075,(B4:0.10,B5:0.10)
    :0.075):0.008577):0.002145,(((C1:0.10,(C2:0.05,C3:0.05):0.05):0.075,(C4
    :0.10,C5:0.10):0.075):0.008577,((D1:0.10,(D2:0.05,D3:0.05):0.05):0.075,(D4
    :0.10,D5:0.10):0.075):0.008577):0.002145);';
```

The largest branches are 0.10, and the smallest branches are the two branches of length 0.002145 (giving the tree its name, Tree 0.002), leading from the bifurcating root node. These two branches are coloured red in H&R Figure 1a, and these red branches each have two descendant clades — (A,B) and (C,D). With reference to their Figure 1a and 1b, inaccuracy was measured by the inability to recover the correct branching pattern of these four daughter clades (and for this part of the analysis, regardless of whether the branching pattern within those clades was correct).

Simulations were made by H&R on this tree using a model derived from the Chang dataset (Chang et al. 2015), including four-category discrete gamma-distributed among-site rate variation with $\alpha = 0.5$, making alignments of 1000 sites. Varied amounts of compositional heterogeneity were imparted to two non-sister clades (clades A and C, H&R Figure 1a). This makes the alignments phylogenetically challenging due to compositional attraction. Analysis was done with RAxML (Stamatakis 2014).

We repeated their analysis and obtained similar results (Table S1). The recoded simulations had worse accuracy compared to unrecoded, as was seen by H&R.

Table S1: Repeat of H&R Figure S2 for Dayhoff analysis of unrecoded data, and Dayhoff recoding. Results are expressed as inaccuracy, the number of analyses out of a total of 1000 that failed to reconstruct the arrangement of the four clades in the simulation tree. The inflation parameter imparts compositional heterogeneity to two non-sister clades; explained in H&R. The simulations made alignments of 1000 sites. For the results in the third row, the first 370 sites of these alignments were re-analysed, unrecoded.

	inflation parameter			
	0.0	0.1	0.5	0.9
unrecoded, all sites	137	139	158	294
Dayhoff recoded	386	372	347	421
unrecoded, 370 sites	487	488	495	612

In these analyses, accuracy is limited by the substitutions between the tips of the two "red" branches. These tips are separated by a total branch length of 0.00429 over the 1000 site simulated alignments. We would therefore expect only 4.29 substitutions on this split. We can look at the realized number of substitutions on the tips of these red branches by simulating data with the same model on a one-branch tree, with branch length of 0.00429. We did 100,000 such simulations, and measured an average of 4.26 unrecoded substitutions, and 1.58 recoded substitutions. The recoded sequences therefore have about 37% of the number of substitutions compared to the unrecoded sequences. H&R explained the poor accuracy of the recoded analyses by the decrease in phylogenetic information of the recoding process. We can test this by repeating the analysis with the first 370 of the 1000 sites of the unrecoded protein alignments. These short alignments would have about the same amount of information as the recoded datasets, and their accuracy is worse than both the unrecoded and the recoded alignments (Table S1). This shows that if we control for the amount of information, and compare recoded accuracy with shortened unrecoded accuracy, then recoding is more accurate in this case.

This suggests that if there was enough information, by a combination of branch length and alignment length, such that enough valid phylogenetic information remained after recoding, then the results would be different. To test this we ask whether the accuracy pattern seen with Tree 0.002 remains if the sequences are more diverged, and so the analysis was repeated with all the branches multiplied ten-fold. The results are shown in Table S2, where recoded analysis is more accurate than unrecoded, reversing the pattern seen in the shorter tree. While the conclusions in Hernandez and Ryan (2021) concerning recoding in the face of compositional heterogeneity over the tree are valid when information is limited, the effect of recoding (increased accuracy in this case) is more clearly apparent when the sequences are more diverged and the amount of information is not as limited.

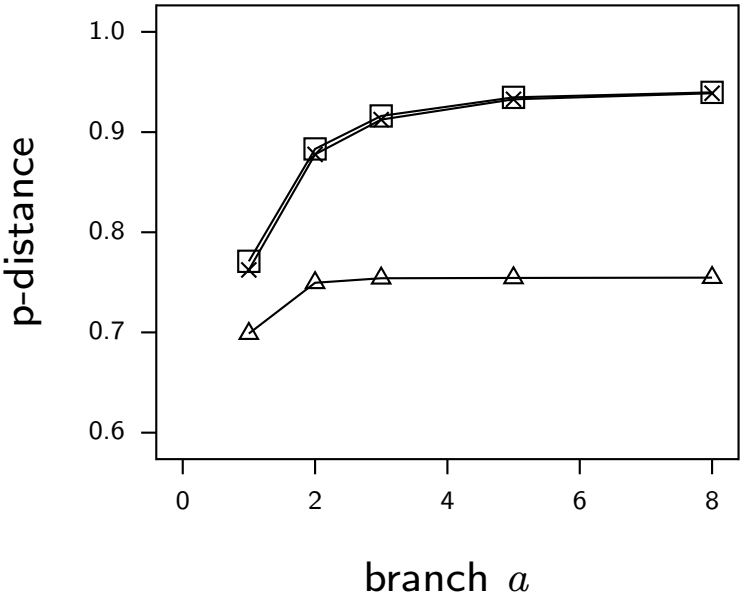
Table S2: Analysis as in Table S1, repeated here with the branch lengths of the simulation tree increased ten-fold. Results are shown as inaccuracy counts out of 1000 replicates.

	inflation parameter			
	0.0	0.1	0.5	0.9
unrecoded	599	568	585	583
Dayhoff recoded	438	446	481	427

Recoding of alignments approaching saturation

Saturation plots

Figure S1: Saturation plots. P-distances were measured between all six leaf pairs of the alignments approaching saturation. The means over 100 alignments are plotted here *vs* the branch a simulation distance as shown in Figure 1 a. Flattening curves at increasing branch length reflect approaching saturation. Values from alignments from Dayhoff78 simulations are shown with $\times$, from LG simulations with a square, and from EDM alignments with a triangle.



Accuracy

Table S3: Accuracy of analyses of simulations approaching saturation. Results show recovery counts of the trees shown in Figure 1 d. The branch a value below refers to the simulation tree in Figure 1 a. Asterisks indicate significant differences ($P < 0.05$) of proportions of Tree I compared to unrecoded analyses. These results are also shown in Figure 2.

	Branch a	Recoding	Tree I	Tree II	Tree III
D78	2.0	None	100	0	0
		KGB	100	0	0
		Dayhoff	100	0	0
		SR-sat	100	0	0
		SR-Chi-sq	99	1	0
	3.0	None	98	2	0
		KGB	97	2	1
		Dayhoff	98	2	0
		SR-sat	98	1	1
		SR-Chi-sq	74*	16	10
	5.0	None	68	19	13
		KGB	62	21	17
		Dayhoff	63	21	16
		SR-sat	57	25	18
		SR-Chi-sq	47*	26	27
	8.0	None	49	22	29
		KGB	38	23	39
		Dayhoff	49	19	32
		SR-sat	47	29	24
		SR-Chi-sq	38	33	29
LG	2.0	None	100	0	0
		KGB	100	0	0
		Dayhoff	100	0	0
		SR-sat	100	0	0
		SR-Chi-sq	96*	2	2
	3.0	None	99	0	1
		KGB	97	0	3
		Dayhoff	98	1	1
		SR-sat	97	2	1
		SR-Chi-sq	67*	15	18
	5.0	None	51	31	18
		KGB	49	29	22
		Dayhoff	55	26	19
		SR-sat	54	31	15
		SR-Chi-sq	43	36	21
	8.0	None	36	32	32
		KGB	34	32	34
		Dayhoff	30	36	34
		SR-sat	38	31	31
		SR-Chi-sq	35	35	30
EDM	2.0	None	67	18	15
		KGB	60	18	22
		Dayhoff	56	19	25
		SR-sat	54	25	21
		SR-Chi-sq	47*	28	25
	3.0	None	39	34	27
		KGB	36	36	28
		Dayhoff	35	38	27
		SR-sat	40	38	22
		SR-Chi-sq	35	36	29
	5.0	None	33	26	41
		KGB	31	28	41
		Dayhoff	32	29	39
		SR-sat	31	33	36
		SR-Chi-sq	31	33	36
	8.0	None	40	32	28
		KGB	39	35	26
		Dayhoff	34	36	30
		SR-sat	32	34	34
		SR-Chi-sq	35	31	34

Chi-squared measurements

Table S4: Chi-squared probabilities for near-saturation simulations and recodings. Simulations were made as described in Figure 1a with accuracy analysis in Figure 2. Chi-squared measurements are shown from alignments with leaf branches of 3 mutations per site.

Model	recoding	min	max	mean	stdev
D78	unrecoded	0.011	0.999	0.594	0.269
	KGB	0.010	0.998	0.597	0.289
	Dayhoff	0.044	0.997	0.587	0.272
	SR-sat	0.011	0.998	0.600	0.279
	SR-Chi-sq	1.000	1.000	1.000	0.000
LG	unrecoded	0.060	0.995	0.570	0.270
	KGB	0.006	0.986	0.580	0.270
	Dayhoff	0.023	0.999	0.568	0.294
	SR-sat	0.011	0.998	0.547	0.273
	SR-Chi-sq	1.000	1.000	1.000	0.000
EDM	unrecoded	0.016	0.999	0.792	0.232
	KGB	0.035	0.999	0.788	0.209
	Dayhoff	0.046	1.000	0.781	0.218
	SR-sat	0.047	0.999	0.789	0.224
	SR-Chi-sq	1.000	1.000	1.000	0.000

Proportion of constant sites

Table S5: Constant sites in near-saturation simulations and recodings. Constant site proportions were measured for the simulated alignments and recodings from Figure 2 for the simulations with branch $a = 3$. Recoding increased the proportion of constant sites, but SR-Chi-sq recoding increased the proportion the least.

	D78		LG		EDM	
recoding	mean	std dev	mean	std dev	mean	std dev
None	0.0037	0.0002	0.0019	0.0001	0.0612	0.0007
KGB	0.0540	0.0032	0.0860	0.0157	0.2785	0.0141
Dayhoff	0.0397	0.0006	0.0412	0.0006	0.2969	0.0015
SR-sat	0.0259	0.0004	0.0347	0.0006	0.2714	0.0015
SR-Chi-sq	0.0202	0.0066	0.0173	0.0069	0.1155	0.0298

Mean site entropy

Table S6: Mean site entropy in near-saturation simulations and recodings. Site entropy values were measured for the simulated alignments and recodings from Figure 2 for the simulations with branch $a = 3$. Means of mean entropy and variance over 100 alignments are shown. Recoding decreased the entropy, but SR-Chi-sq decreased the entropy the least, and also had the smallest variance of the recodings. EDM entropy was smaller, but with larger variance, compared to the single-matrix D78 and LG values.

	D78		LG		EDM	
recoding	mean	var.	mean	var	mean	var
None	1.7538	0.1154	1.7611	0.1054	1.3818	0.2970
KGB	1.1892	0.2157	1.1166	0.2461	0.7984	0.3353
Dayhoff	1.2728	0.2085	1.2698	0.2105	0.7899	0.3560
SR-sat	1.3432	0.1912	1.3103	0.2058	0.8352	0.3614
SR-Chi-sq	1.4052	0.1836	1.4140	0.1767	1.1073	0.2911

Variance of sitewise character frequency

Table S7: Variance of sitewise character frequency in near-saturation simulations and recodings. Values were measured for the simulated alignments and recodings from Figure 2 for the simulations with branch $a = 3$. Means of variance over 100 alignments are shown. Variances of EDM simulations and recodings had bigger variances than those for the site-homogeneous D78 and LG models. Of the recodings, SR-Chi-sq had the smallest variance.

Recoding	D78	LG	EDM
None	0.0128	0.0127	0.0186
KGB	0.0368	0.0366	0.0640
Dayhoff	0.0386	0.0398	0.0691
SR-sat	0.0393	0.0401	0.0693
SR-Chi-sq	0.0364	0.0360	0.0533

Specific recodings

Table S8: Specific recoding groups obtained for the near-saturation series of simulations. The recodings for the first ten alignments of the three simulation models are shown for KGB recoding and for SR-Chi-sq recoding. SR-Chi-sq recoding shows great variability compared to KGB recoding — in the 30 SR-Chi-sq recodings shown there are 172 different groups, while in the 30 KGB recodings there are only 21 different groups.

Simulation	KGB groups	SR-Chi-sq groups
D78	rndqehkt, cv, agps, ilm, w, fy	netw, aqly, mfps, cgv, rhi, dk
	rndqehk, cv, agpst, ilm, w, fy	hkmy, dwv, rlps, anci, gf, qet
	rndqehkt, cv, agps, ilm, w, fy	aglfw, npyv, ds, qehk, cmt, ri
	rndqehkt, cv, agps, ilm, w, fy	nyv, lp, dk, acqmstw, rehi, gf
	rndqehk, cv, agpst, ilm, w, fy	adp, wy, ceift, rnlk, qhv, gms
	rndqehkt, cv, agps, ilm, w, fy	rcf, nitv, msy, egkp, qhw, adl
	rndqehkt, cv, agps, ilm, w, fy	ev, iy, qft, arglkms, dc, nhpw
	rndqehkt, cv, agps, ilm, w, fy	gf, anks, ipt, rdqh, clmv, ewy
	rndqehk, cv, agpst, ilm, w, fy	qf, mstwy, eil, apv, rdgk, nch
	rndqehkt, cs, agpv, ilm, fy, w	dc, ks, egmfyv, alw, hipt, rnq
LG	andegpst, rc, qhk, ilmv, fy, w	kp, lfswy, rc, aqit, ndh, egmv
	andegpst, rc, qhk, ilmv, fy, w	ceif, gty, rlsw, dhm, anpv, qk
	andegpst, rc, qhk, ilmv, w, fy	adk, cqiwv, nehlf, rgs, t, mpy
	andegpst, rc, qhk, ilmv, fy, w	iv, nqhty, rcfp, es, dglkm, aw
	andegpst, rc, qhk, ilmv, fy, w	diy, fsw, arkv, nemt, cq, ghlp
	andegpst, rc, qhk, ilmv, fy, w	swyv, ilp, dft, anc, rgk, qehm
	andegpst, rc, qhk, ilmv, fy, w	eyv, rmstw, dp, cgi, nqlf, ahk
	andegpst, rc, qhk, ilmv, fy, w	rcpv, dty, ahw, qgil, nks, emf
	rnqehkp, adgst, c, ilmv, w, fy	dpt, nms, aqv, clky, rehiw, gf
	rnqehkp, adgst, c, ilmv, fy, w	lpt, qimw, degfv, c, arky, nhs
EDM	cs, andegpt, rqhk, ilmv, fy, w	qhkv, cswy, g, ndept, ai, rlmf
	cs, andegpt, rqhk, ilmv, fy, w	cqv, ry, defs, ak, gilmtw, nhp
	cs, angptv, rdqehk, ilm, fy, w	ncq, ekmpw, hif, dg, stv, arly
	cs, andegpt, rqhk, ilmv, fy, w	chikf, arstwv, qeg, py, dl, nm
	cs, andegpt, rqhk, ilmv, fy, w	cqw, nemv, aips, dhkt, rgly, f
	cs, andegpt, rqhk, ilmv, fy, w	hw, qgy, acs, dt, nlmfp, reikv
	cs, andegpt, rqhk, ilmv, fy, w	t, emyv, cip, ardls, nkf, qghw
	cs, andegpt, rqhk, ilmv, fy, w	niv, qsty, egk, archl, dfw, mp
	cs, andegpt, rqhk, ilmv, fy, w	efwv, cqt, ds, ahpy, nik, rglm
	cs, andegpt, rqhk, ilmv, fy, w	nqeft, yv, rilkp, amsw, ch, dg

Recoding of alignments with compositional heterogeneity over the tree

Accuracy

Table S9: Accuracy of analysis of simulations with compositional heterogeneity over the tree. Results show accuracy as recovery counts of Tree I shown in Figure 1 d, out of 100 simulations, with increasing branch c of the simulation tree. Analysis used the LG+F model (the simulation model with free composition) for unrecoded alignments, and the six-state GTR model for recoded alignments. Tree II was never recovered. Recovery from Series A (compositionally homogeneous control in Table 1) was Tree I for all simulations.

Simulation Series [a]	Recoding	0.05	0.10	0.20
B	None	0	99	100
	KGB	40	100	100
	Dayhoff	100	100	100
	SR-sat	85	100	100
	SR-Chi-sq	1	41	100
C	None	0	0	35
	KGB	0	25	100
	Dayhoff	0	0	100
	SR-sat	0	0	100
	SR-Chi-sq	2	88	100
D	None	0	9	100
	KGB	0	77	100
	Dayhoff	0	100	100
	SR-sat	0	0	100
	SR-Chi-sq	0	1	61

[a] Compositional heterogeneity imparted to the simulations on the b-branches of the simulation tree of Figure 1 b, referring to the simulation series in Table 1. The proportions of the affected amino acids of the simulation model were adjusted up or down by a factor of ten, and then normalized.

Accuracy with among-site rate variation

Table S10: Accuracy of analysis of simulations with compositional heterogeneity over the tree, simulated with gamma-distributed among-site rate variation (ASRV) with four categories and the shape parameter $\alpha = 0.5$. This is a repeat of the analysis shown in Figure 3 a but with ASRV. Results show accuracy as recovery counts of Tree I shown in Figure 1 d, out of 100 simulations, with increasing branch c of the simulation tree. Analysis used the LG+F+G4 model (the simulation model with free composition) for unrecoded alignments, and the six-state GTR+G4 model for recoded alignments. Tree II was never recovered. Recovery from Series A (compositionally homogeneous control in Table 1) was Tree I for all simulations. Simulation Series B, C, D indicates compositional heterogeneity imparted to the simulations on the b-branches of the simulation tree of Figure 1 b, referring to the simulation series in Table 1. The proportions of the affected amino acids of the simulation model were adjusted up or down by a factor of ten, and then normalized.

Simulation Series	Recoding	Length of internal branch c 0.05	0.10	0.20	0.25	0.30
B	None	0	0	100	100	100
	KGB	59	98	100	100	100
	Dayhoff	99	100	100	100	100
	SR-sat	1	100	100	100	100
	SR-Chi-sq	6	45	100	100	100
C	None	0	0	0	0	84
	KGB	0	0	0	0	0
	Dayhoff	0	0	0	0	0
	SR-sat	0	0	1	80	100
	SR-Chi-sq	1	77	100	100	100
D	None	0	0	99	100	100
	KGB	0	0	99	100	100
	Dayhoff	0	1	100	100	100
	SR-sat	0	0	6	98	100
	SR-Chi-sq	0	4	94	100	100

Mean site entropy

Table S11: Mean site entropy in compositionally tree-heterogeneous simulations. Site entropy values were measured for the simulated alignments and recodings from Figure 3 for the simulations with branch $c = 0.05$. Means of mean entropy and variance over 100 alignments are shown. Recoding decreased the entropy, although SR-Chi-sq decreased the entropy the least, and (except for series A) also had the smallest variance of the recodings.

	Series A		Series B		Series C		Series D	
recoding	mean	var	mean	var	mean	var	mean	var
None	1.4097	0.2693	1.3850	0.2711	1.3632	0.2847	1.4583	0.2535
KGB	0.5504	0.2689	0.7533	0.2882	0.6825	0.2822	0.6267	0.2770
Dayhoff	0.8437	0.3111	0.8181	0.3066	0.7916	0.2862	0.9300	0.3016
SR-sat	0.8825	0.3158	0.8518	0.3127	0.8779	0.2908	0.9669	0.3015
SR-Chi-sq	1.1456	0.2714	1.1483	0.2696	0.9353	0.2526	1.2098	0.2627

Variance of sitewise character frequency

Table S12: Variance of sitewise character frequency in the compositionally tree-heterogeneous simulations. Values were measured for the simulated alignments and recodings from Figure 3 for the simulations with branch $c = 0.05$. Means of the variance over 100 alignments are shown. Of the recodings, SR-Chi-sq had the smallest variance.

Recoding	A	B	C	D
None	0.0182	0.0186	0.0184	0.0168
KGB	0.0576	0.0641	0.0549	0.0503
Dayhoff	0.0675	0.0688	0.0646	0.0602
SR-sat	0.0676	0.0690	0.0651	0.0601
SR-Chi-sq	0.0527	0.0536	0.0483	0.0499

Specific recodings

Table S13: KGB and SR-Chi-sq recodings in compositionally tree-heterogeneous results shown in Figure 3. Recoding schemes for the first ten alignments are shown. Compared to KGB recoding, SR-Chi-sq recoding varies widely over the replicate alignments, as shown by the number of unique groups in the ten sample recodings.

series	KGB recoding	SR-Chi-sq recoding	SR-Chi-sq n unique groups
Control	arndcqehkst g ilmfv p w y	eipt lwyv ang kmf rcqh ds	60
	arndcqehkst g ilmfv p w y	p dcqft lsv rnim gwy aehk	
	arndcqehkst g ilmfv p w y	hmt qgpw rnciv elfy d aks	
	arndcqehkst g ilmfv p w y	qsty clv rnf dhip akw egm	
	arndcqehkst g ilmfv p w y	rekp itwy dhfv anclm s qg	
	arndcqehkst g ilmfv p w y	rmfwv gly ep acit q ndhks	
	arndcqehkst g ilmfv p w y	ar lmstw kpyv ncg def qhi	
	arndcqehkst g ilmfv p w y	dhtv rqf nim ekw acs glpy	
	arndcqehkst g ilmfv p w y	yv nhsw rkfp ilt aqegm dc	
	arndcqehkst g ilmfv p w y	acqtv ef ndghil rms k pwy	
Dayhoff	arncqekst dg ilmv fy hp w	aps hfv rny cetw qgik dlm	44
	andcgst rqek ilmv hp w fy	nqtv dew cil kpsy armf gh	
	arndcqekst g ilmv fy hp w	efsv kty ndlp aqi gmw rch	
	andcgst rqek ilmv hp w fy	dqy hlt swv ceik armf ngp	
	andcgst rqek ilmv hp w fy	rdpy elst hfv nck gmw aqi	
	andcgst ilmv rqek hp w fy	npwy elst dchv rqf agi km	
	andcgst rqek ilmv hp w fy	hy dmpv lkfs cetw rng aqi	
	rnqekst adcg ilmv hp w fy	hy dqlt csv eikw amfp rng	
	arncqekst dg ilmv fy hp w	swv nqety ckf dlm gh arip	
	andcgst rqek ilmv hp w fy	hfv elst cgm dqy nkw arip	
GAP-FYMINK	andceghstv rqk ilm fy p w	gifty w h qemp arndcksv l	30
	andceghstv rqk ilm fy p w	w arncikmfv gsty h l dqep	
	andceghstv rqk ilm fy p w	gks andqeifty rcmpv l w h	
	andceghstv rqk ilm fy p w	l akmstyv w h cqefp rndgi	
	andceghstv rqk ilm fy p w	w cfpyv rdegi anqkmst h l	
	andceghstv rqk ilm fy p w	l areiktyv cgmfs h ndqp w	
	andceghstv rqk ilm fy p w	l cgmfs areiktyv h ndqp w	
	andceghstv rqk ilm fy p w	w cqkp dgtyv arneimfs l h	
	andceghstv rqk ilm fy p w	gimyv adqekst l rncfp w h	
	andceghstv rqk ilm fy p w	cmfpv ardqeikt l ngsy w h	
Arbitrary	arndcqehkst g ilmfv p w y	kmy nftv asw egip rch dql	38
	arndcqehkst g ilmfv p w y	ewv aimps qfty rch nk dgl	
	arndcqehkst g ilmfv p w y	ctv hmps aiky nfw rqe dgl	
	arndcqehkst g ilmfv p w y	mftyv esw rch nk dilp aqg	
	arndcqehkst g ilmfv p w y	neis mpwyv dt lkf aqg rch	
	arndcqehkst g ilmfv p w y	npwv eist rch aqf kmy dgl	
	arndcqehkst g ilmfv p w y	ewv qfty agps dilm nk rch	
	arndcqehkst g ilmfv p w y	lfpsw neiv hky aqg dt rcm	
	arndcqehkst g ilmfv p w y	astv ref qhpw dgl cimy nk	
	arndcqehkst g ilmfv p w y	ewv gpsty rch dilm aqf nk	

Accuracy of DNA recodings

Table S14: Accuracy of recoded compositionally tree-heterogeneous DNA alignments. Simulation series are described in Table 4. Unrecoded alignments were analysed with the F81 model, and recoded alignments were analysed with a 2-state model, both with free composition parameters. The MK recoding scheme recoded A and C as one character, and recoded G and T as another, RY recoded A and G as one character and C and T as another, and WS recoded A and T as one character and C and G as another. These results are plotted in Figure 4. The Chi-squared values of the alignments were measured, and mean P-values over all alignments are shown.

Simulation Series	Recoding	Tree I	Tree II	Tree III	mean P_{χ^2}
E	None	100	0	0	0.48
	MK	98	1	1	0.49
	RY	99	0	1	0.50
	WS	98	0	2	0.52
F	None	0	0	100	0.0
	MK	0	0	100	0.0
	RY	99	0	1	0.50
	WS	0	0	100	0.0
G	None	0	0	100	0.0
	MK	99	1	0	0.54
	RY	14	0	86	0.0
	WS	6	0	94	0.0
H	None	0	0	100	0.0
	MK	0	0	100	0.0
	RY	0	0	100	0.0
	WS	0	0	100	0.0

Recoding of alignments that are compositionally heterogeneous over alignment sites

Accuracy

Table S15: Effect of recoding on analysis of compositionally site-heterogeneous alignments. Accuracy is shown as counts, out of 100 simulations, of recovery of Tree I in Figure 1 d, over increasing branch a values in the simulation tree. Tree II was never recovered. VT+F was used for unrecoded analyses (recoding None) and for making the KGB recoding. Recoded data analysis used a 6-state GTR model.

Recoding	0.10	0.15	0.20
None	0	0	100
KGB	0	0	0
Dayhoff	0	0	0
SR-sat	0	0	3
SR-Chi-sq	0	24	99

Simulation and analysis with among-site rate variation

Table S16: Effect of recoding on analysis of compositionally site-heterogeneous alignments that were simulated with discrete gamma distributed among-site rate variation (ASRV) with 100 categories and the shape parameter $\alpha = 0.5$. Accuracy is shown as counts, out of 100 simulations, of recovery of Tree I in Figure 1 d, over increasing branch a values in the simulation tree. Tree II was never recovered. VT+F+G4 was used for unrecoded analyses (recoding None) and for making the KGB recoding. Recoded data analysis used a 6-state GTR+G4 model. These results can be compared to those in Figure 5 a, which was the same analysis but without ASRV. Results here are similar to those without ASRV in that unrecoded analysis and SR-Chi-sq recoding were best, while the exchangeability-based recodings KGB, Dayhoff, and SR-sat were less accurate.

Recoding	0.10	0.15	0.20
None	0	0	100
KGB	0	0	37
Dayhoff	0	0	74
SR-sat	0	0	82
SR-Chi-sq	0	0	100

Character diversity

Table S17: Character diversity (CD) in compositionally site-heterogeneous alignments. Control alignments were simulated with the single-matrix LG model, and site-heterogeneous simulations used the EDM model. Measurements were made on alignments from trees shown in Figure 1 c with an internal branch length of 0.15, with accuracy shown in Figure 5 a

recoding		mean CD	stdev	range
unrecoded	control	2.408	0.003	2.398 – 2.415
unrecoded	EDM	2.280	0.003	2.273 – 2.288
KGB	control	1.466	0.002	1.461 – 1.469
KGB	EDM	1.486	0.005	1.481 – 1.536
Dayhoff	control	1.743	0.002	1.738 – 1.748
Dayhoff	EDM	1.706	0.002	1.701 – 1.712
SR-sat	control	1.782	0.002	1.777 – 1.786
SR-sat	EDM	1.750	0.002	1.744 – 1.756
SR-Chi-sq	control	2.097	0.059	1.937 – 2.234
SR-Chi-sq	EDM	2.017	0.053	1.809 – 2.100

Chi-squared measurements

Table S18: Chi-squared measurements from compositionally site-heterogeneous alignments. Control alignments were simulated with the single-matrix LG model, and site-heterogeneous alignments were simulated with the EDM model. Measurements were made on alignments from trees shown in Figure 1 c with an internal branch length of 0.15, for which reconstruction accuracy is shown in Figure 5 a

recoding		mean Chi-sq	stdev	range	mean P	stdev	range
unrecoded	control	36.490	7.401	22.602 – 61.236	0.946	0.109	0.327 – 1.000
unrecoded	EDM	34.602	6.093	17.842 – 48.545	0.972	0.050	0.780 – 1.000
KGB	control	6.469	2.591	1.693 – 14.630	0.938	0.101	0.478 – 1.000
KGB	EDM	7.769	2.717	2.788 – 17.209	0.896	0.125	0.307 – 1.000
Dayhoff	control	7.394	3.018	2.631 – 16.809	0.903	0.143	0.330 – 1.000
Dayhoff	EDM	7.234	2.927	2.124 – 22.478	0.916	0.132	0.096 – 1.000
SR-sat	control	7.148	2.864	2.139 – 16.650	0.913	0.128	0.340 – 1.000
SR-sat	EDM	7.332	2.781	3.017 – 20.233	0.912	0.126	0.163 – 1.000
SR-Chi-sq	control	0.195	0.090	0.070 – 0.568	1.000	0.000	1.000 – 1.000
SR-Chi-sq	EDM	0.189	0.074	0.045 – 0.447	1.000	0.000	1.000 – 1.000

Mean site entropy

Table S19: Mean site entropy in compositionally site-heterogeneous simulations. Control alignments were simulated with the single-matrix LG model, and site-heterogeneous alignments were simulated with the EDM model. Measurements were made on alignments from trees shown in Figure 1 c with an internal branch length of 0.15, for which reconstruction accuracy is shown in Figure 5 a. Means of mean entropy and variance over 100 alignments are shown. Recoding lowered the mean site entropy relative to unrecoded entropy, although SR-Chi-sq lowered it the least.

	Control LG		EDM	
recoding	mean	var	mean	var
None	1.0704	0.3219	0.9889	0.3186
KGB	0.3835	0.2208	0.4046	0.2344
Dayhoff	0.5983	0.2844	0.5737	0.2955
SR-sat	0.6272	0.2921	0.6068	0.3028
SR-Chi-sq	0.8664	0.2977	0.8115	0.2954

Variance of sitewise character frequency

Table S20: Variance of sitewise character frequency in compositionally site-heterogeneous simulations. Control alignments were simulated with the single-matrix LG model, and site-heterogeneous alignments were simulated with the EDM model. Measurements were made on alignments from trees shown in Figure 1 c with an internal branch length of 0.15, for which reconstruction accuracy is shown in Figure 5 a. Means of the variance over 100 alignments are shown. Of the recodings, SR-Chi-sq had the smallest variance.

Recoding	LG	EDM
None	0.0242	0.0255
KGB	0.0699	0.0695
Dayhoff	0.0846	0.0837
SR-sat	0.0852	0.0846
SR-Chi-sq	0.0699	0.0728

Specific recodings

Table S21: Specific KGB and SR-Chi-sq recodings in compositionally site-heterogeneous alignments. Single-matrix compositionally site-homogeneous LG simulations are shown as a control. Compositionally site-heterogeneous simulations used the EDM model and were made as described in Figure 1 c with an internal branch length of 0.15, with analysis results in Figure 5 a. One hundred replicate simulations were made for each; the first ten grouping schemes are shown.

Simulation	KGB recoding	SR-Chi-sq recoding
LG	arndcqehkst g ilmfv p w y	nqks cmt giv efp hwy ardl
	arndcqehkst g ilmfv p w y	rcm sty dev nqghik afw lp
	arndcqehkst g ilmfv p w y	dmfpyv lkst nw ar qgi ceh
	arndcqehkst g ilmfv p w y	emv qfp rckt dil nswy agh
	arndcqehkst g ilmfv p w y	celt dkps nqhyv gi af rmw
	arndcqehkst g ilmfv p w y	gmps adft qek rchl wy niv
	arndcqehkst g ilmfv p w y	rgfv qeimpw hlsy ad c nkt
	arndcqehkst g ilmfv p w y	sv ikfwy acqp ndg t rehlm
	arndcqehkst g ilmfv p w y	gftv mp anksy c rhi dqelw
	arndcqehkst g ilmfv p w y	aksw rcmt ndly giv qh efp
EDM	arndcqehkst g ilmfv p w y	lm nfwv ikpsy dq ach regt
	arndcqehkst g ilmfv p w y	hity npsv re dcgm qlf akw
	arndcqehkst g ilmfv p w y	ght ly qfpv rcsw ekm andi
	arndcqehkst g ilmfv p w y	v ghlms de ariw qfpy nckt
	arndcqehkst g ilmfv p w y	anhlksy im rpt ewv df cqg
	arndcqehkst g ilmfv p w y	istyv rlw acg nmf qh dekp
	arndcqehkst g ilmfv p w y	empy ndcgst ifw alk rhv q
	arndcqehkst g ilmfv p w y	defv lmtw nqps acgk rh iy
	arndcqehkst g ilmfv p w y	achms lfyv qk dw gip rnet
	arndcqehkst g ilmfv p w y	det cwyv aghiks rqf nm lp

Recoding of alignments that are both site- and tree-heterogeneous

Accuracy

Table S22: Accuracy of analysis of simulations with compositional heterogeneity both over the alignment sites and over the tree. Accuracy is shown at increasing branch a lengths as counts of recovery of Tree I in Figure 1 d. Tree II was never recovered. For unrecoded alignments, analysis used the VT+F model as chosen using ModelFinder in IQ-TREE. Analysis of recoded alignments used the GTR model.

Simulation Series [a]	Recoding	0.20	0.25	0.30	0.35	0.40
B	None	0	0	34	100	100
	KGB	0	4	97	100	100
	Dayhoff	0	4	100	100	100
	SR-sat	0	2	100	100	100
	SR-Chi-sq	0	1	29	77	98
C	None	0	40	100	100	100
	KGB	0	0	39	100	100
	Dayhoff	0	6	100	100	100
	SR-sat	0	18	100	100	100
	SR-Chi-sq	0	61	100	100	100
D	None	0	0	0	57	100
	KGB	0	0	0	25	91
	Dayhoff	0	0	60	100	100
	SR-sat	0	0	2	92	100
	SR-Chi-sq	0	0	2	33	84

[a] Compositional heterogeneity imparted to the simulations on the b-branches of Figure 1 c, referring to the simulation series in Table 1. The proportions of the affected amino acids of the simulation model were adjusted up or down by a factor of three, and then normalized.

Chi-sqared measurements

Table S23: Compositional heterogeneity of alignments from Figure 6. Chi-squared measurements were taken from alignments from simulations with an internal branch length of 0.30.

	Series B		Series C		Series D	
Recoding	Chi-squared	P	Chi-squared	P	Chi-squared	P
None [a]	33808 – 35100	0.0	13275 – 14090	0.0	34811 – 36208	0.0
KGB	2253 – 2780	0.0	2278 – 3417	0.0	1610 – 3594	0.0
Dayhoff	597 – 818	0.0	5036 – 5592	0.0	1967 – 2285	0.0
SR-sat	3777 – 4289	0.0	3422 – 3910	0.0	5584 – 6046	0.0
SR-Chi-sq	1.7 – 11.7	1.0 – 0.70	0.5 – 6.1	1.0 – 0.98	2.7 – 14.0	1.0 – 0.52

[a] Unrecoded alignments were 20-state, while recoded alignments were 6-state, making the Chi-squared measurements not directly comparable.

Mean site entropy

Table S24: Mean site entropy of alignments that are compositionally both site- and tree-heterogeneous. The alignments from simulations with an internal branch length of 0.30 were measured. Means of mean entropy and variance over 100 alignments are shown. Accuracy is shown in Figure 6. Recoding lowered the entropy relative to unrecoded entropy, although SR-Chi-sq lowered it the least.

	Series B		Series C		Series D	
recoding	mean	var	mean	var	mean	var
None	1.1493	0.3058	1.1084	0.3277	1.1348	0.3183
KGB	0.4520	0.2508	0.6158	0.2949	0.5298	0.2646
Dayhoff	0.6621	0.3149	0.6429	0.3103	0.6692	0.3174
SR-sat	0.6956	0.3220	0.6930	0.3179	0.7059	0.3236
SR-Chi-sq	0.9448	0.2940	0.8906	0.2964	0.9373	0.3028

Variance of sitewise character frequency

Table S25: Variance of sitewise character frequency of alignments that are compositionally both site- and tree-heterogeneous, for which accuracy is shown in Figure 6. The alignments from simulations with an internal branch length of 0.30 were measured. Means of variance over 100 alignments are shown. Of the recodings, SR-Chi-sq had the smallest variance.

Recoding	B	C	D
None	0.0224	0.0230	0.0225
KGB	0.0653	0.0615	0.0588
Dayhoff	0.0775	0.0758	0.0762
SR-sat	0.0779	0.0773	0.0764
SR-Chi-sq	0.0652	0.0637	0.0654

Specific recodings

Table S26: Recoding schemes used in compositionally site- and tree-heterogeneous alignments. Alignments were from simulations made with an internal branch length of 0.30, with accuracy results shown in Figure 6. The first 10 of the 100 replicate alignment groupings are shown for each simulation series.

Simulation	KGB recoding	SR-Chi-sq recoding
B	arndcqehkst g ilmfv p w y	qepsy mfv ahtw rcil nk dg
	arndcqehkst g ilmv p w fy	ety qiv dg ckf apsw rnhlm
	arndcqehkst g ilmv fy p w	qemfy htv apsw nk ril dcg
	arndcqehkst g ilmv fy p w	hpwy dmv nlt rgks cef aqi
	arndcqehkst g ilmfv p w y	gm swv rdk ety aqi nchlfp
	arndcqehkst g ilmfv p w y	cqmp rhy aitwv elfs nk dg
	arndcqehkst g ilmfv p w y	nqp htv cilw akfsy em rdg
	arndcqehkst g ilmv fy p w	gptw ndv elfs aqi ckm rhy
	arndcqehkst g ilmv fy p w	swv efp dqgh ily rckt anm
	arndcqehkst g ilmfv p w y	de iwyv rqmp ckf ahlst ng
C	arndqehkmst cgv il fy p w	dgfs lwy aeikm rq nptv ch
	arndqehkmst cgv il fy p w	anikv eghmty qw dfp rc ls
	arndqehkmst cgv il fy p w	gks qhw adeimty np rc lfv
	arndqehkmst cgv il fy p w	andekfy mpt ls cgiwv r qh
	arndqehkmst cgv il fy p w	fp qtwv aikm negs rdly ch
	arndqehkmst cgv il fy p w	fp ly ahikst dev rcq ngmw
	arndqehkmst cgv il fy p w	etv lwy adiks ncp gmf rqh
	arndqehkmst cgv il fy p w	ls ancif ehmp dtwv rq gky
	arndqehkmst cgv il fy p w	tv qls rdhw cgmf np aeiky
	arndqehkmst cgv il fy p w	qtv egms ailkfy np rc dhw
D	arndcqehkmstv g ilf p w y	nmv ekfy pst adgw qil rch
	arndcqehkmstv g ilf p w y	qy efw dlkmps arci htv ng
	rndqehkst acg ilmv fy p w	ptv ekmf nsy qil adgw rch
	arndcqehkmstv g ilf p w y	dch astv fpwy eg nim rqlk
	rndqehkst acg ilmv fy p w	qm ilftyv nkw adg eps rch
	rndqehkst acg ilmv fy p w	ehs ptv niy ckmw agf rdql
	arndcqehkmstv g ilf p w y	nsy himp efw ctv adg rqlk
	arndcqehkmstv g ilf p w y	qls degy ptv akfw nim rch
	arndcqehkmstv g ilf p w y	qm eyv ilk nfpst argw dch
	arndcqehkmstv g ilf p w y	hmps eyv cit dqlk nf argw

Analysis using heterogeneous models

Tree-heterogeneous alignments: Accuracy using NDCH

Table S27: Simulations with compositional heterogeneity over the tree as in Figure 3 a but analysed with the NDCH model. Results show accuracy as recovery counts of Tree I shown in Figure 1 d, out of 100 simulations. Tree II was never recovered. Recovery from Series A (compositionally homogeneous control in Table 1) was Tree I for all simulations.

		analysis with homogeneous models [1]			analysis with NDCH model		
Series	Recoding	0.05	0.10	0.20	0.05	0.10	0.20
B	None	0	99	100	100	100	100
	KGB	40	100	100	100	100	100
	Dayhoff	100	100	100	100	100	100
	SR-sat	85	100	100	100	100	100
	SR-Chi-sq	1	41	100	1	38	100
C	None	0	0	35	100	100	100
	KGB	0	25	100	99	100	100
	Dayhoff	0	0	100	100	100	100
	SR-sat	0	0	100	97	100	100
	SR-Chi-sq	2	88	100	1	71	100
D	None	0	9	100	100	100	100
	KGB	0	77	100	99	100	100
	Dayhoff	0	100	100	97	100	100
	SR-sat	0	0	100	99	100	100
	SR-Chi-sq	0	1	61	0	1	57

[1] Results from Figure 3 a are repeated here.

Site-heterogenous alignments: Accuracy using the CAT model

Table S28: Effect of recoding on analysis of compositionally site-heterogeneous alignments using the CAT model. Accuracy is shown as counts, out of 100 simulations, of recovery of Tree I in Figure 1 d. Tree II was never recovered.

	analysis with homogeneous models [a]			analysis with CAT model		
Recoding	0.10	0.15	0.20	0.10	0.15	0.20
None	0	0	100	100	100	100
KGB	0	0	0	3	41	100
Dayhoff	0	0	0	16	98	100
SR-sat	0	0	3	17	99	100
SR-Chi-sq	0	24	99	100	100	100

[a] Results from Figure 5 a are repeated here.

Site- and tree-heterogeneous alignments: Accuracy using NDCH and CAT

Table S29: Analysis of simulations with compositional heterogeneity both over the alignment sites and over the tree. Accuracy is shown at increasing branch a lengths as counts of recovery of Tree I in Figure 1 d. Tree II was never recovered.

Analysis model	Simulation Series	Recoding	0.20	0.25	0.30	0.35	0.40
homogeneous [a]	B	None	0	0	34	100	100
		KGB	0	4	97	100	100
		Dayhoff	0	4	100	100	100
		SR-sat	0	2	100	100	100
		SR-Chi-sq	0	1	29	77	98
	C	None	0	40	100	100	100
		KGB	0	0	39	100	100
		Dayhoff	0	6	100	100	100
		SR-sat	0	18	100	100	100
		SR-Chi-sq	0	61	100	100	100
	D	None	0	0	0	57	100
		KGB	0	0	0	25	91
		Dayhoff	0	0	60	100	100
		SR-sat	0	0	2	92	100
		SR-Chi-sq	0	0	2	33	84
NDCH	B	None	100	100	100	100	100
		KGB	0	63	100	100	100
		Dayhoff	0	7	100	100	100
		SR-sat	0	98	100	100	100
		SR-Chi-sq	0	0	21	73	98
	C	None	100	100	100	100	100
		KGB	0	11	100	100	100
		Dayhoff	30	100	100	100	100
		SR-sat	1	100	100	100	100
		SR-Chi-sq	0	57	100	100	100
	D	None	100	100	100	100	100
		KGB	0	0	3	78	100
		Dayhoff	0	0	99	100	100
		SR-sat	0	22	100	100	100
		SR-Chi-sq	0	0	2	31	81
CAT	B	None	42	66	96	100	99
		KGB	9	51	100	100	100
		Dayhoff	98	100	100	100	100
		SR-sat	24	99	100	100	100
		SR-Chi-sq	78	90	95	98	98
	C	None	100	100	100	100	100
		KGB	9	43	100	100	100
		Dayhoff	8	45	100	100	100
		SR-sat	44	100	100	100	100
		SR-Chi-sq	96	100	100	100	100
	D	None	11	10	18	63	100
		KGB	7	12	36	66	99
		Dayhoff	60	99	100	100	100
		SR-sat	11	21	72	99	100
		SR-Chi-sq	30	54	68	87	99

[a] Results from Figure 6 a are repeated here.

Using SR-Chi-sq to choose the number of bins

Based on personal communication from Ed Susko and Andrew Roger, the authors of the SR-Chi-sq strategy and software, the analysis shown in Figure 3 a was repeated such that an additional recoding strategy was used where minmax-chisq results were used to choose the number of bins. In other analyses in this study, SR-Chi-sq has been used by specifying 6 bins. The additional strategy (SR-Chi-sq-iterate) is to start with 19 bins and iterate down to smaller bin sizes. Again, we used 10 random starts for each trialled number of bins, and again set minmax-chisq to use 10000 rearrangements. The iteration stops when one of the P-values from the minmax-chisq software is greater than 0.05, and the bins when that happens were chosen (the highest of the P-values from the ten starts was used).
The datasets from Figure 3 were re-analysed this way. The number of bins chosen (Table S30) depended on the simulation series (Dayhoff, GAP-FYMINK, or Arbitrary, Table 1). Phylogenetic accuracy is shown in Table S31. Accuracy of SR-Chi-sq-iterate appeared to be somewhat better than SR-Chi-sq for the Dayhoff-series simulations, although not as accurate as the exchangeability-based recodings. For simulation series GAP-FYMINK and Arbitrary, accuracy of SR-Chi-sq-iterate did not differ greatly from SR-Chi-sq.

Table S30: Average number of bins found by the SR-Chi-sq-iterate binning strategy

Simulation Series	0.05	0.10	0.20
Dayhoff	7.55	7.52	7.47
GAP-FYMINK	4.91	4.94	4.93
Arbitrary	6.84	6.84	6.83

Table S31: Accuracy of alignments from Figure 3 re-analysed to include the SR-Chi-sq-iterate binning strategy. SR-Chi-sq results differed slightly from those in Figure 3 a due to use of different random number seeds.

Simulation Series	Recoding	0.05	0.10	0.20
Dayhoff	None	0	99	100
	KGB	40	100	100
	Dayhoff	100	100	100
	SR-sat	85	100	100
	SR-Chi-sq	1	37	100
	SR-Chi-sq-iterate	5	91	100
GAP-FYMINK	None	0	0	35
	KGB	0	25	100
	Dayhoff	0	0	100
	SR-sat	0	0	100
	SR-Chi-sq	1	88	100
	SR-Chi-sq-iterate	0	82	100
Arbitrary	None	0	9	100
	KGB	0	77	100
	Dayhoff	0	100	100
	SR-sat	0	0	100
	SR-Chi-sq	0	2	65
	SR-Chi-sq-iterate	0	4	69

References

Chang, E. S., M. Neuhof, N. D. Rubinstein, A. Diamant, H. Philippe, D. Huchon, and P. Cartwright (2015) Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. *Proc Nat Acad Sci (USA)* 112, 14912–14917. DOI: 10.1073/pnas.1511468112.

Hernandez, A. M. and J. F. Ryan (2021) Six-state amino acid recoding is not an effective strategy to offset compositional heterogeneity and saturation in phylogenetic analyses. *Syst Biol*, 1200–1212. DOI: 10.1093/sysbio/syab027.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. DOI: 10.1093/bioinformatics/btu033.