Online Resource 1

Journal of Molecular Evolution

The standard genetic code's rank among randomized alternatives is not a property of the code

Lucas Giovani Ribeiro, Marcos André Simonssini

Independent Researchers, Uruana de Minas, Minas Gerais, Brazil

Corresponding author: Lucas Giovani Ribeiro — coin26@mail.com

Contents

Supplementary Note S1 — Protocol adherence and deviations

The analysis protocol was frozen on 27 July 2026, before any confirmatory analysis was executed, and is

deposited in full with its complete changelog. This note records every departure from it during execution.

All are implementation-level; none alters a prospectively specified hypothesis, endpoint, estimand, metric

or exclusion rule.

Before any species-level spectrum was loaded

D1 — Normalization wording and implementation. The protocol text specified renormalization "per codon and position". The first implementation took this literally and rescaled each codon–position block to equal mass, which forces every codon position to contribute equally and suppresses the relative weighting of positions. Detected by an additivity check on the decomposition and corrected to a single global sense→sense normalization.

D2 — Synthetic calibration superseded. Calibration figures first computed without context integration were inflated approximately two-fold, because the CpG multiplier had been applied to all C>T substitutions rather than only to C>T in genuine CpG context. All calibration figures were recomputed under the final pipeline; the superseded values appear nowhere in the analysis.

D3 — Confirmatory metric lock. A synthetic sensitivity sweep, run before any species-level spectrum was loaded, showed that the effect size depends on the amino-acid metric. The confirmatory metric was locked to squared polar requirement on historical-standard grounds, a criterion independent of effect size, and the lock was not revised after the cross-metric analysis showed the effect to be largest under that metric. The metric dependence is reported in the main text.

D4 — Cohort order. A polymorphism-based, pentanucleotide-resolution cohort was designated the principal study in the protocol. Its deposit is a single 4.3 GB archive that could not be retrieved as a whole, and the present cohort was executed first for reasons of data access alone.

Posterior-summary files for 68 of its 113 species were recovered later, in an order fixed before any of their spectra were read, and all 68 were analysed; none was examined and then excluded. All 113 species of the deposit are listed in data/external_source_manifest.csv with their position in its alphabetical order and whether each was recovered, so the selection can be checked rather than taken on trust. Because the cohort remained incomplete it is reported as an external replication rather than as the principal confirmatory set. This was recorded in the protocol before any result existed. The present cohort is reported standalone, with the resulting limitation stated in the main text, and is not presented as though it had always been the principal study.

After species-level data were loaded

D5 — Out-of-fold prediction. Fixed effects for each fold were estimated by inverse-variance-weighted generalised least squares on the training clades rather than by full phylogenetic refitting within each fold.

This is a computational approximation to the specified procedure.

D6 — Per-clade diagnostic slopes. Within-clade slopes were first computed by ordinary least squares as a diagnostic. Under the phylogenetic model the fungal CpG coefficient — which under OLS appeared an order of magnitude larger than in any other clade — collapses to approximately zero. The interpretation built on the uncorrected estimate was withdrawn, and only phylogenetically corrected estimates are reported.

D7 — Unresolved taxon names. TimeTree returned 58 tips under labels absent from the dataset,

representing synonym substitutions. In keeping with the protocol's requirement that unresolved names be

listed and excluded rather than silently substituted, these were dropped rather than matched by

approximation. Exclusion lists are deposited.

Supplementary Note S2 — Analysis protocol and its revision history

The protocol was frozen as version 5 on 27 July 2026, before the confirmatory analyses were run. It

specifies the observable and its estimand, the confirmatory amino-acid metric, the randomized-code

ensemble and its checksum, the aggregation rule, the treatment of stop codons and strand collapsing, the

hypotheses and their directional predictions, the cross-validation folds, and the exclusion rules — each

fixed before the data were seen.

It carries a complete internal changelog, dated through thirteen earlier revisions,

recording every decision that was made, reversed, or reopened during development,

and why. Two reversals are worth noting because they changed the estimand rather

than the implementation:

• The context exposure was first specified as each species' own coding-sequence

composition and was reversed to a fixed reference exposure common to all species. Species-specific

exposure would have meant the observable moved because of the spectrum, codon usage and contextual

composition jointly, rather than the spectrum alone.

• The test of the primary null was rebuilt. The original per-species bootstrap

resampled each species around its own observed spectrum and therefore preserved the between-species

differences it was meant to test; it was replaced by a measurement-error phylogenetic model evaluated by

parametric bootstrap under the null.

Protocol and changelog: see Supplementary Note S3.

Supplementary Table S1 — Effect of transition bias within each clade

Maximum-likelihood estimates of the coefficient of log transition/transversion ratio on the spectrum-

conditioned advantage, fitted with the phylogenetic model separately within each clade. Intervals are 95%

profile-likelihood.

Clade n β_κ 95% CI λ

Insects 489 +0.0697 +0.0624 to +0.0770 0.996

Clade n β_κ 95% CI λ

Other 212 +0.0567 +0.0446 to +0.0688 0.998

Mammals 669 +0.0478 +0.0425 to +0.0532 0.990

Fungi 619 +0.0474 +0.0423 to +0.0525 0.997

Plants 903 +0.0460 +0.0409 to +0.0511 0.993

Birds 897 +0.0411 +0.0365 to +0.0456 0.985

Reptiles 134 +0.0337 +0.0261 to +0.0414 1.000

Fish 1,049 +0.0320 +0.0285 to +0.0356 0.992

All eight estimates are positive and all eight intervals exclude zero. The coefficient varies by a factor of 2.2 across the set.

Supplementary Table S2 — Out-of-clade prediction error, decomposed

Leave-one-clade-out validation. RMSE for the phylogeny-only model (M0) and the spectrum model (M1); the M1 error is decomposed into an intercept-bias component and a dispersion component.

Held-out n RMSE M0 RMSE M1 bias (M1) dispersion bias share of

clade (M1) MSE

Mammals 669 0.01875 0.00579 +0.00120 0.00567 4.3%

Reptiles 134 0.01007 0.00886 −0.00613 0.00640 47.9%

Birds 897 0.01542 0.00929 +0.00278 0.00887 8.9%

Fish 1,049 0.03103 0.01223 −0.00747 0.00968 37.3%

Plants 903 0.03337 0.01439 −0.00508 0.01347 12.5%

Insects 489 0.04211 0.01762 −0.01089 0.01385 38.2%

Fungi 619 0.02295 0.02196 +0.0000009 0.02196 0.0%

Other 212 0.04587 0.02685 −0.00987 0.02496 13.5%

Weighted overall: RMSE 0.02877 (M0) to 0.01458 (M1), a gain of 49.3%. After recalibrating each fold's intercept: 0.01726 to 0.01333, a gain of 22.7%.

Supplementary Table S3 — Inverted leave-one-clade-out validation

The model is trained on a single clade and used to predict all remaining species. Recalibrated gain removes the intercept offset.

Training n train n test β_κ RMSE M0 RMSE M1 gain recalibrated

clade

Mammals 669 4,303 +0.01672 0.03468 0.01611 53.6% 39.7%

Training n train n test β_κ RMSE M0 RMSE M1 gain recalibrated

clade

Fish 1,049 3,923 +0.02198 0.02890 0.01481 48.7% 42.3%

Reptiles 134 4,838 +0.01831 0.02805 0.01495 46.7% 39.9%

Plants 903 4,069 +0.01975 0.02705 0.01467 45.8% 40.5%

Birds 897 4,075 +0.01288 0.03461 0.02194 36.6% 33.8%

Other 212 4,760 +0.01360 0.03301 0.02180 34.0% 28.8%

Fungi 619 4,353 +0.01441 0.02982 0.03023 −1.4% −5.8%

Insects 489 4,483 +0.01749 0.03260 0.05414 −66.1% −3.5%

All eight training clades yield a positive coefficient, spanning a factor of 1.7. The spectrum model outperforms the phylogeny-only model on the rest of the tree from six of eight. The insect shortfall is almost entirely a level offset.

Supplementary Table S4 — Cross-metric comparison

The spectrum-conditioned advantage recomputed under three amino-acid distance metrics, and the fitted

effect of transition bias scaled by the dispersion of the advantage under that metric.

Metric mean R sd R range of R β_κ β_κ / sd(R)

Squared Woese 0.5477 0.0246 26.46 pp +0.0206 0.836

polar requirement

(confirmatory)

Grantham 0.2018 0.0077 10.61 pp +0.0041 0.529

Squared Kyte– 0.3345 0.0304 44.30 pp +0.0120 0.395

Doolittle

All values are computed over the 4,972 species of the confirmatory set. β_κ is the coefficient on

standardized log κ from an ordinary least-squares fit that also includes the standardized CpG>T shift;

deposited as data/cross_metric.csv.

The association is strongest under squared polar requirement and weaker under two alternative metrics.

These are correlated with polar requirement rather than independent of it: the scales at −0.79, their

squared-difference matrices at +0.43. The ordering does not triangulate polarity as the mediating axis, and

no sign reversal occurs.

Supplementary Table S5 — In-silico code optimization

Simulated annealing over permutations of the twenty amino acids across the twenty standard codon

blocks, under spectra with fixed transition/transversion ratio and no CpG term. Forty restarts of 30,000

iterations per value. Similarity is the fraction of blocks receiving the same amino acid as the standard

code; chance level is 1/20 = 5%.

κ Φ optimum Φ standard similarity

1.0 3.489 5.194 12.4%

1.5 3.488 4.890 7.9%

2.0 3.280 4.663 3.1%

3.0 2.879 4.346 4.1%

4.0 2.593 4.136 3.8%

6.0 2.174 3.874 2.1%

8.0 1.893 3.717 4.5%

Correlation of similarity against log κ: −0.76.

Supplementary Table S6 — Stop-codon placement, exact enumeration

Nonsense burden depends only on which codons are stops and on the spectrum, not on the assignment of

amino acids to the remaining codons. All C(64,3) = 41,664 three-codon placements were therefore

enumerated exactly under the reference exposure and the mean spectrum.

quantity value

burden of the standard set {UAA, UAG, UGA} 0.03789

distribution across all placements min 0.02374, median 0.04736, max 0.08032

percentile of the standard set, all placements 10.8

placements with lower burden 4,514 of 41,664

percentile among placements of equal pairwise Hamming 37.7

spread (n = 1,728)

Clustering, not the particular choice of codons, accounts for the apparent advantage. Correlation between

pairwise spread and burden across all placements is +0.20.

Supplementary Table S7 — Sensitivity: species with a non-standard nuclear code

Sixty-four species in the confirmatory set belong to the CUG clade, which reassigns CUG from leucine to

serine. No ciliates or other variant-code lineages were present.

set n mean R sd R range β_κ corr(κ, R)

full 4,972 0.5477 0.0246 26.46 pp 0.05815 0.831

excluding 4,908 0.5478 0.0246 26.46 pp 0.05827 0.838

CUG clade

Supplementary Table S8 — Position of the standard code within the ensemble

Under the synthetic sweep, the standard code is compared both against the frozen 10,000-code ensemble

and against the optimum located by simulated annealing.

κ Φ standard best ensemble codes Φ annealed optimum /

code outperforming optimum standard

SGC

1 5.194 5.111 3 of 10,000 3.489 0.672

2 4.663 4.663 0 of 10,000 3.280 0.703

4 4.136 4.218 0 of 10,000 2.593 0.627

8 3.717 3.864 0 of 10,000 1.893 0.509

16 3.439 3.501 0 of 10,000 — —

The ratio of the best ensemble code to the standard code rises from 0.984 to 1.040 while the ratio of the

annealed optimum to the standard code falls from 0.672 to 0.509. A change of scale would move both in

the same direction.

Supplementary Table S9 — Exclusion of RIP as the fungal residual

Within-clade correlation between log κ and R, with fungal species partitioned by the presence of repeat-

induced point mutation.

Group n corr(κ, R) sd(R)

Pezizomycotina (RIP 369 +0.440 0.0152

present)

Yeasts (no RIP) 58 +0.617 0.0269

Remaining fungi 192 +0.416 0.0287

Basidiomycota (no RIP) 47 +0.462 0.0153

Other clades, for comparison — +0.54 to +0.86 —

Groups lacking RIP show the same shortfall as those possessing it, so the mechanism does not account for

the fungal residual.

Supplementary Table S10 — The CUG reassignment under yeast spectra

Expected substitution cost under the standard code and under the CUG-clade code (CUG reassigned from

leucine to serine), evaluated for each species under its own spectrum and the same reference exposure. Δ

is the CUG-code cost minus the standard-code cost; negative values favour the reassignment.

Group n mean Δ median Δ % with Δ < 0

CUG clade (carries the 64 −0.0186 +0.0042 43.8%

reassignment)

Other yeasts (standard 86 −0.0363 −0.0016 51.2%

code)

Non-yeast fungi 469 +0.0801 +0.0787 1.1%

Mann–Whitney between the CUG clade and other yeasts: p = 0.18. The contrast with non-yeast fungi

separates yeasts from non-yeasts rather than carriers from non-carriers, and reflects phylogenetic structure

in the spectra.

Supplementary Table S11 — Slope and level of the standard code within the ensemble

The regression of R on log κ repeated for all 10,000 randomized codes against the same species spectra.

quantity standard code ensemble median percentile of standard code

β_κ (per unit of log κ) +0.0582 +0.0148 70.5

mean R 0.5477 0.0219 99.98

A positive coefficient occurs in 5,676 of the 10,000 codes. Responsiveness to transition bias is therefore

largely a property of the block structure; the level of advantage is not.

Supplementary Table S12 — Curvature in the out-of-clade intercept bias

Leave-one-clade-out validation repeated with higher-order terms in log κ, weighted across folds by fold

size.

model RMSE weighted bias dispersion gain over bias share of

`intercept-only MSE`

intercept only 0.02877 — — — —

linear 0.01458 0.00591 0.01333 49.3% 16%

(confirmatory)

plus κ^2 0.01434 0.00416 0.01372 50.2% 8%

plus κ^2 and κ^3 0.01466 0.00418 0.01405 49.1% 8%

Out-of-fold intercept bias by clade, against clade mean log κ:

Clade mean log κ bias (linear) bias (plus κ^2)

Insects 0.570 +0.01089 +0.00617

Other 0.618 +0.00988 +0.00755

Plants 0.641 +0.00508 +0.00410

Fish 0.748 +0.00747 +0.00502

Fungi 1.110 −0.00000 −0.00089

Birds 1.155 −0.00278 −0.00279

Reptiles 1.162 +0.00613 +0.00619

Mammals 1.210 −0.00120 −0.00137

Correlation between clade mean log κ and bias: −0.82 under the linear model. The near-zero bias for

Fungi is a crossing point of the linear approximation rather than a property of the clade: it moves by three

orders of magnitude once the quadratic term is added, while the biases of high-κ clades barely change.

Supplementary Table S13 — Which stop codon a nonsense mutation creates

Sense→stop exposure partitioned by the stop codon generated, and by whether the substitution is a

transition or a transversion. "Routes" counts codon–position– context combinations rather than distinct

codon pairs: positions 1 and 3 each open into four trinucleotide contexts and position 2 into one, so the 23

distinct sense→stop codon pairs expand to 71 context-resolved routes. Exposure is the summed reference

weight of those routes.

Stop codon routes transition exposure transversion tv/ts

exposure

UAA 22 1.0 6.0 6.00

UAG 23 2.0 6.0 3.00

UGA 26 2.0 6.0 3.00

The two transition routes to UGA are CGA→UGA (C>U at position 1) and UGG→UGA (G>A at

position 3).

Share of nonsense burden by stop codon, across quintiles of κ:

κ quintile mean κ UAA UAG UGA

lowest 1.52 0.263 0.295 0.442

second 1.96 0.231 0.280 0.489

third 2.45 0.207 0.262 0.531

fourth 3.08 0.179 0.244 0.577

κ quintile mean κ UAA UAG UGA

highest 4.04 0.152 0.216 0.632

Correlation with log κ: UAA −0.77, UAG −0.51, UGA +0.67. The transversion share of total nonsense

burden falls from 0.54 in the lowest quintile to 0.26 in the highest (correlation −0.94), so transition bias

makes nonsense both smaller and more transition-carried.

Supplementary Table S14 — Worst-case position of the standard code

For each species, the number of randomized codes in the frozen 10,000-code ensemble achieving a lower

expected cost than the standard code under that species' own spectrum.

quantity value

median codes outperforming the standard code 2 of 10,000

worst observed spectrum *Plasmodium gaboni*, 341 of 10,000

implied worst-case position 342 of 10,001, i.e. the top 3.5%

species within the top 0.5% of the ensemble 4,969 of 4,972 (99.94%)

species where no randomized code outperforms it 1,084 of 4,972 (21.8%)

Mean advantage across the 4,972 spectra places the standard code third of 10,001, against a median

ensemble mean of 0.0219. This is close to the roughly two-in-ten-thousand reported by Haig and Hurst

(1991) under uniform mutation, so the classical ranking is not materially altered by substituting empirical

spectra.

We do not read the accompanying variance as evidence of exceptional stability. The standard code sits at

the 10.3rd percentile of across-species variance, but mean and variance are correlated at −0.52 across the

ensemble, and among the five codes with a comparable mean it falls at the 20th percentile of variance.

With only two codes exceeding it on the mean, the absence of one that also undercuts it on variance would

arise about 80% of the time under independence. All statements here are confined to the block-preserving

ensemble.

Supplementary Note S3 — Reproduction package

All analysis code, the frozen protocol with its internal dated changelog, the randomized-code

ensemble, the taxon-name mapping, the exclusion lists and all derived per-species values are provided as

Online Resource 2 (ESM_2.zip), published with this article. The package is self-contained with respect to

the deposited code, derived outputs and the numerical checks it runs. Full reconstruction from the primary

mutation-spectrum data, and five check groups comprising twelve checks, including the within-clade

correlations, the nonsense-channel quantities and the curvature checks, require the third-party CORAL

supplementary table cited in the main text.

The package includes three verification utilities. verify_package.py uses only the Python standard library

and confirms the file manifest, the SHA-256 checksum of the frozen ensemble

(d35d209a380b3a9db11ac809f0daa120c0ee91b9dbb23299c910aa2e906763e0), its structural invariants,

the row counts of five principal derived tables and the tip count of the phylogeny. audit_numbers.py recomputes

every quantity asserted in the manuscript from the deposited outputs and compares it to the value reported;

it reports 150 checks from the deposited outputs, or 162 when the external branch-level table is supplied.

Substitution counts and callable abundances aggregated per species are included so the audit runs without

external downloads; the branch-level source table itself is not redistributed and is obtained from the

supplementary material of the source study cited in the main text.

Supplementary Note S4 — Stability of the phylogenetic fit

Three diagnostics, regenerated by code/phylo_diagnostics.py, on all 4,972

confirmatory species. The join uses the binomial column, since the tree is

keyed by binomial while the spectra carry strain and subspecies suffixes.

Likelihood profile. Profiling lambda with tau-squared maximized at each point

gives a maximum at lambda = 0.9922 and a 95% profile interval of 0.9911 to

0.9931. Profiling tau-squared with lambda optimized continuously at each point

gives 3.49e-03 to 4.02e-03 around a maximum of 3.74e-03, which reproduces the

estimate of the main model. Intervals are read from the

curves rather than from an asymptotic approximation, because tau-squared = 0

lies on the boundary of the parameter space and lambda is unidentified there.

The profiles are shown in Figure S1.

Simulation recovery. Data were generated on the same tree at the fitted

tau-squared and at lambda = 0.50, 0.90 and 0.99, then refitted by the same

procedure. Six replicates at each value:

```
lambda generated recovered, median range tau-squared ratio
0.50 0.529 0.463 to 0.569 1.06
0.90 0.902 0.897 to 0.906 1.02
0.99 0.990 0.990 to 0.990 1.01
```

A lambda near one is therefore not an artefact of the estimator saturating at

this sample size and with this measurement-error structure: when the generating

value is 0.50 the estimator returns 0.50.

Polytomy resolution order. The covariance is built by sequential pairwise

merging, which is exact under Brownian motion only if the order in which

polytomies are resolved does not affect the likelihood. Resolving them six times

at random gives lambda = 0.99217 in every case,

and a coefficient of variation in tau-squared of 1.2e-16. That is floating-point

noise, not a dependence on order.

Supplementary Note S5 — Convergence of the in-silico optimization

Simulated annealing cannot prove it has found a global optimum, so this note

records what the budget achieved rather than asserting what it ruled out.

Regenerated by code/optim_convergence.py.

Three budgets at each of three transition/transversion ratios:

```
kappa restarts iterations best cost similarity restarts at best
```

```
1.0 10 30,000 3.4886 0.10 2 of 10
1.0 40 30,000 3.4886 0.10 12 of 40
1.0 40 120,000 3.4886 0.10 25 of 40
2.0 10 30,000 3.2795 0.05 7 of 10
2.0 40 30,000 3.2795 0.05 14 of 40
2.0 40 120,000 3.2795 0.05 38 of 40
4.0 10 30,000 2.5925 0.05 2 of 10
4.0 40 30,000 2.5925 0.05 13 of 40
4.0 40 120,000 2.5925 0.05 22 of 40
```

Raising the total budget sixteen-fold changes the best cost by 0.00 per cent at every kappa tested, and the similarity at the optimum does not move. What rises is the fraction of independent restarts that reach the best cost, from 7 of 10 to 38 of 40 at kappa = 2. Restarts are converging on the same value from independent starting points, which is the evidence available for a heuristic search: not a proof of global optimality, but a demonstration that the reported optima are reproducible and that more computation does not find better ones. The similarity at the optimum is 5 to 10 per cent throughout, at or below the level expected by chance for a permutation of twenty blocks.

Supplementary Note S6 — Strand canonicalization

The source spectra collapse reverse-complementary substitution types. Applying the collapsed rate to both members of a complementary pair requires flipping the context as well as the bases; copying the numeric rate alone is incorrect. The lookup rule is central base C or T → query the context as presented central base A or G → query reverse_complement(context) with the complementary substitution P_s(G → A | x) = P_s(C → T | RC(x)) (7) Collapsed spectra assume strand symmetry and therefore cannot represent transcription-coupled-repair asymmetry in coding regions (§12).

Supplementary Table S15 — Advantage and exceptionalness under transition bias, by clade

Coefficients are per unit of log κ, controlling for the CpG shift. Intervals are 95% profile likelihood. Per-species values are deposited as data/cv_z_by_clade.csv.

clade	n	β CV	95% CI	β z	95% CI	p (z)
Birds	897	+0.0417	+0.0394 to +0.0439	-0.2449	-0.2770 to -0.2128	< 1e-15
Fish	1049	+0.0446	+0.0426 to +0.0465	-0.3133	-0.3347 to -0.2920	< 1e-15
Fungi	619	+0.0147	+0.0123 to +0.0171	-0.0080	-0.0436 to +0.0276	0.658
Insects	489	+0.0365	+0.0336 to +0.0394	-0.2280	-0.2619 to -0.1941	< 1e-15
Mammals	669	+0.0405	+0.0375 to +0.0434	-0.2106	-0.2444 to -0.1767	< 1e-15
Other	212	+0.0253	+0.0204 to +0.0301	-0.0571	-0.1606 to +0.0464	0.278
Plants	903	+0.0376	+0.0354 to +0.0397	-0.2962	-0.3390 to -0.2533	< 1e-15
Reptiles	134	+0.0495	+0.0440 to +0.0550	-0.2988	-0.3527 to -0.2449	< 1e-15

The coefficient for CV is positive in all eight clades, with every interval excluding zero. The coefficient for z is negative in all eight by sign, and its interval excludes zero in six: Fungi and the pooled remainder are the exceptions, and both are groups in which κ accounts for less of the variation in R generally.

Supplementary Figure S1

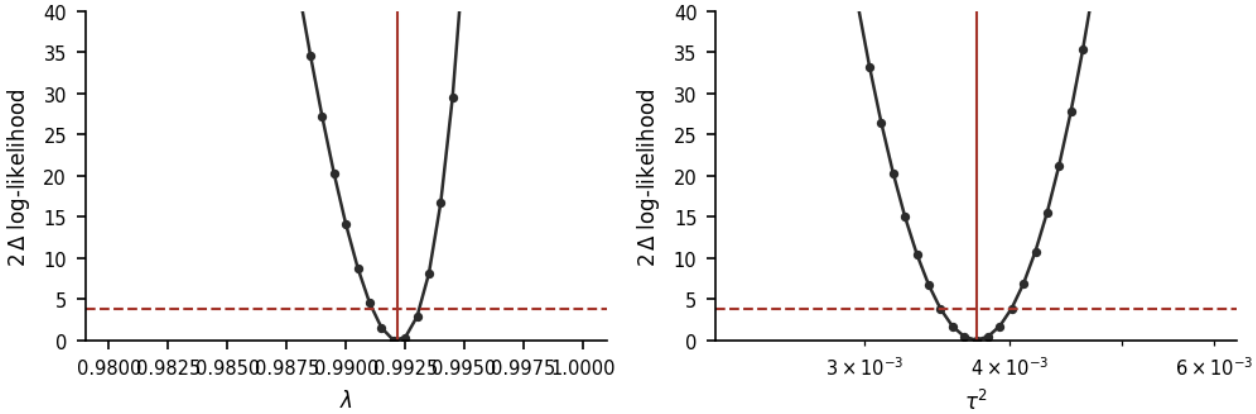

Fig. S1 Likelihood profiles for the phylogenetic model over all 4,972 confirmatory species. Left: 2 delta log-likelihood against lambda, with tau-squared maximized at each point. Right: the same against tau-squared, with lambda optimized continuously at each point. The dashed line marks 3.84, the vertical line the maximum-likelihood estimate.

Supplementary Figure S2

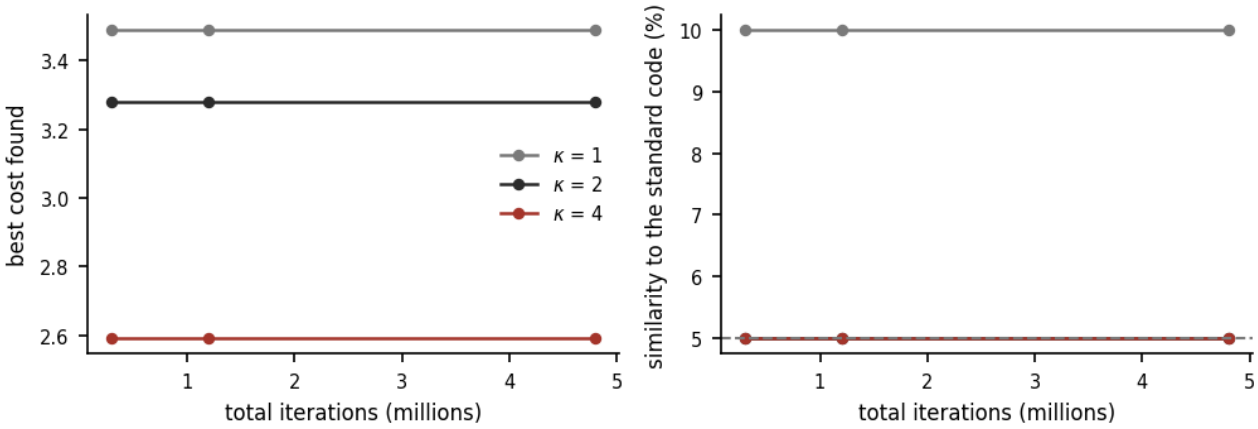

Fig. S2 The in-silico optimization does not improve with a larger budget. Left: best cost found against total annealing iterations, at three transition/transversion ratios. Right: similarity of the best solution to the standard code, with the dashed line at the 5% expected by chance.