Supplementary material for "Exact Bayesian evidence for phylogenetic trees"

Matthew D. Schwartz*
Department of Physics, Harvard University

Scott V. Edwards
Department of Organismic and Evolutionary Biology, Harvard University

Paul O. Lewis
Department of Ecology and Evolutionary Biology, University of Connecticut

Contents

*Work done with Anthropic, but results and views are not endorsed by Anthropic.

S1 Beyond the rational case: rate variation and tempering

Section 3.4 of the main text states that under rate variation across sites and under tempered likelihoods the term-by-term argument no longer applies, and the marginal likelihood is then a constant of another kind, presumably irrational. This section gives the two smallest cases in full: the constants that appear under the $+\Gamma$ model, and the linear differential equation that determines a tempered evidence at $\beta = \frac{1}{2}$. The vocabulary of periods and differential operators used here does not appear in the main text and is not needed to read it.

S1.1 Rate variation across sites

Under the $+\Gamma$ model of Yang (1993) each site carries a rate $r \sim \mathrm{Gamma}(k, 1/k)$. Integrating the rate out turns each Jukes–Cantor pattern probability into a sum of terms $(1 + \frac{1}{k}\sum_{e \in S} t_e)^{-k}$ over edge subsets S. These terms couple the edges, the evidence integral no longer factorizes

edge by edge, and the term-by-term route to a fraction is closed. On the three-taxon star tree with shape $k = 1$, an exponential branch-length prior and a single constant site,

$$Z_{+\Gamma} \;=\; \frac{10 - 6\,G}{64}, \qquad G \;:=\; e\,E_1(1) \;=\; \int_0^\infty \frac{e^{-t}}{1+t}\,\mathrm{d}t \;=\; 0.59634736\ldots \tag{S1}$$

where G is the Euler–Gompertz constant, whose irrationality is an open problem. The closed form agrees with an independent one-dimensional tanh–sinh quadrature to 120 digits, and the constant G itself was checked to 220. In the classification of Kontsevich and Zagier (2001) this is an exponential period, an integral of e^{-f} times an algebraic function, rather than a period of an algebraic variety.

At two or more sites a second constant enters, $I_2 = \int_1^\infty E_1(u)\,u^{-1}\mathrm{d}u$. An integer-relation search (PSLQ) at 160 digits finds I_2 independent of $\{1, G, G^2, \gamma, \pi^2, \log 2\}$, and the set of constants grows with the alignment length as a tower of iterated exponential integrals. The governing differential equation is the order-2 confluent (Kummer) equation satisfied by $e^{\lambda}E_1(\lambda)$, which has an irregular singular point of rank one at infinity. Near an irregular point the solutions involve genuine exponentials of the parameter, not only powers and logarithms, and the standard tools for evaluating solutions of Fuchsian equations with rigorous error bounds assume the latter. The $+\Gamma$ evidence therefore lies outside the class of functions to which the transport method of Section S1.2 applies as it stands.

S1.2 The tempered evidence at one half

A stepping-stone analysis samples from $L(x)^\beta$ times the prior at non-integer inverse temperatures β (Lartillot and Philippe, 2006; Xie et al., 2011). For non-integer β the integrand $\prod_k p_k(x)^{\beta y_k}$ has non-integer exponents, and the evidence is an Euler integral of Gelfand–Kapranov–Zelevinsky (GKZ) type in the strict sense (Gel'fand et al., 1990; Lee and Pomeransky, 2013). The device is the same as dimensional regularization in the Feynman-integral literature, where integer exponents are shifted off the integers by a continuous parameter. By Hotta's theorem the GKZ system of the integrand's monomial support (its Cayley configuration, which we verified to be homogeneous) is regular holonomic for every exponent vector. The tempered evidence must therefore satisfy a Fuchsian differential equation, one whose singular points are all regular. We establish this concretely on the three-taxon star tree with $\mathbf{y} = \{\mathrm{xxx}{:}2, \mathrm{xxy}{:}1, \mathrm{xyz}{:}1\}$, $\alpha = 1$ and $\beta = \frac{1}{2}$. There $Z(\frac{1}{2}) = 64^{-2}\hat{Z}$ with $\hat{Z} = \int_{[0,1]^3} f_0\sqrt{f_1 f_2}\,\mathrm{d}x$ and f_i the rescaled pattern polynomials.

Geometry. This paragraph classifies the geometry behind the integral. Its one consequence, used by the closed-form search below, is that the constant should lie in the simplest, logarithm-level class of periods, with nothing elliptic or beyond. The double cover $w^2 = f_1 f_2$ is a nodal Fano double cover of $(\mathbb{P}^1)^3$ with $h^{3,0} = 0$. Both branch factors are linear in x_1 with a shared constant part, the x_1-discriminant $16x_2^2(1 - x_3)^2(1 - x_2x_3)^2$ is a perfect square, every slice is a genus-0 conic, and the bundle has a rational section. The threefold is rational with trivial intermediate Jacobian, so the period is expected to be mixed-Tate of weight at most 3: no elliptic curve, K3 surface or Calabi–Yau variety appears anywhere in the pencil.

The differential equation. The object computed is a linear differential equation that the evidence satisfies as a function of a temperature-like deformation parameter t (an annihilating,

or Picard–Fuchs, operator). It plays for the temperature path the role that pruning plays for a single likelihood evaluation: a finite exact rule that reduces every evaluation to mechanical propagation. Here propagation means solving the equation from a point where the value is known exactly, with rigorous error control. Deforming along the coefficient line via $g_i = f_i - 1$,

$$\hat{Z}(t) = \int f_0 \sqrt{(1+tg_1)(1+tg_2)}\,\mathrm{d}x,$$

so that $t = 1$ is the tempered evidence and $\hat{Z}(0) = 4$ exactly, the minimal annihilating operator has order 10 and polynomial degree 33, with leading coefficient

$$q_{10} \;=\; 4\,t^7\,(t-1)^5\,(3t+1)^2\,S_{19}(t), \tag{S2}$$

S_{19} irreducible of degree 19. The operator was found by a nullspace search on the Taylor coefficients of $\hat{Z}(t)$ computed modulo 31-bit primes to order 260 and reconstructed over the integers by the Chinese remainder theorem and rational reconstruction. That it annihilates $\hat{Z}$ is then proved by creative telescoping (Zeilberger, 1990). Write $F = f_0\sqrt{D}$ with $D = (1+tg_1)(1+tg_2)$ for the integrand and $L = \sum_{j=0}^{10} q_j(t)\,\mathrm{d}^j/\mathrm{d}t^j$ for the operator. We exhibit rational functions

$$C_e = \frac{x_e(1-x_e)\,A_e(t,x)}{t^{16}\,f_0\,D^9}, \qquad e = 1,2,3,$$

with A_e polynomials of degree at most 51 in t and at most 16 in each x_i, such that

$$LF = \sum_{e=1}^{3} \frac{\partial}{\partial x_e}\big(C_e F\big) \tag{S3}$$

identically in (t, x). Cleared of denominators, Eq. (S3) is an equality of two integer polynomials in (t, x_1, x_2, x_3) of 70,820 terms each, which we checked by exact expansion; the certificate A_e (110,392 integer coefficients of up to 315 bits) and the checking script are in the supplementary files `certificate_order10.json` and `ct_verify.py`. For $0 < t < 1$ both factors $1 + tg_i = (1-t) + tf_i$ are positive on the closed cube, so differentiation in t commutes with the integral and integrating Eq. (S3) over $[0,1]^3$ gives $L\hat{Z}(t) = \sum_e \int \big[C_e F\big]_{x_e=0}^{x_e=1} = 0$, the boundary terms vanishing through the factor $x_e(1-x_e)$; since $\hat{Z}$ is analytic on $(-\frac{1}{3}, 1)$, $L\hat{Z} = 0$ there. The certificate was obtained by solving the linear system that Eq. (S3) imposes on the coefficients of the A_e, prime by prime with rational reconstruction; modulo two primes, no telescoper of order below 10 exists with certificates of this form. Independently of the proof, the operator annihilates the Taylor coefficients of $\hat{Z}$ recomputed from the definition at a prime used nowhere in its construction (orders 0 to 300) and the exact rational coefficients to order 60 (the released evaluator repeats this check to order 13 before every use). Its transport of the exact initial data at $t = 0$ agrees with independent quadrature at $t = \frac{1}{4}, \frac{3}{10}, \frac{1}{2}$ and 1 (below), and its integer coefficients are given in the supplementary file `tempered_ode_order10.json`. The local exponents (the powers with which solutions behave near each singular point) were computed exactly at every singular point. The nineteen roots of S_{19} are apparent singularities, where the equation degenerates but every solution stays regular (integer exponents $0, \ldots, 8, 10$), and the four true singular points $\{0, -\frac{1}{3}, 1, \infty\}$ are all regular. The operator is therefore Fuchsian, as Hotta's theorem requires. The true singularities are endpoint Landau loci in the sense of Lee and Pomeransky (2013): $t = 1$ is where the branch divisor pinches the vertices of the cube. At

that point the exponents include the quarter-integers $\{\frac{9}{4}, \frac{5}{2}, \frac{11}{4}\}$, an order-4 local monodromy that is invisible in the square-root integrand. The quarter-integers fix the function class of the evidence near $t = 1$: local behavior $(1 - t)^{9/4}$ cannot come from the square roots visible in the integrand, and any closed-form candidate must reproduce it.

The value. Transporting the exact operator from exact rational initial conditions at $t = 0$ and resumming the regular-singular expansion at $t = 1$ gives

$$\hat{Z}(1) \;=\; 2.1466466\ldots, \tag{S4}$$

where the resummation uses Richardson elimination on the exact exponent ladder. That is, successive extrapolations cancel the known subleading powers one at a time. The value agrees to 135 digits with an independent quadrature (the x_3 integral in closed form, panel Gauss–Legendre quadrature in the other two variables in ball arithmetic, so that rounding error is bounded rigorously while the discretization error is controlled by doubling the node count and the working precision), with interior ordinary-point checks at 66 to 80 digits. The digits quoted are those on which the two routes agree; neither route alone supplies a certified enclosure at that precision. The two routes share no method, so an error in either would have to be matched by an identical error in the other to go unseen. The digits are also what the closed-form search below needs, since such a search has no discriminating power at low precision.

Identification of the constant remains open. A two-precision PSLQ search was run against an 18-generator mixed-Tate $\mathbb{Z}[\frac{1}{6}]$ basis through weight 3, a candidate list of known logarithm-level constants. At integer-coefficient heights to 10^5 with 146 digits, and through weight 2 at heights to 10^9, it returns nothing stable, with positive controls passing. Symbolic iterated integration through the explicit conic rationalization splits the constant into three pieces, $\hat{Z}(1) = P_{\mathrm{rat}} + P_{\mathrm{sqrt}} + P_{\log}$. The rational piece has the closed form

$$P_{\mathrm{rat}} = -\tfrac{851}{144} + \tfrac{1555}{48}\log 2 - \tfrac{17}{512}\pi^2 - \tfrac{1925}{64}\log 2 \log 3 + \tfrac{1925}{128}\,\mathrm{Li}_2(-2) - \tfrac{1925}{256}\,\mathrm{Li}_2(-8),$$

matched by an independent contour computation to 251 digits (whose two precision levels agree to 210). The square-root piece closes in a pure $\sqrt{3}$/sixth-root class of constants, confirmed by an independent route to 135 digits (its two precision levels agree to 125). The logarithmic piece reduces to a one-dimensional integral $\int_0^1 T(w)\,dw$ with T rational in w against letters of weight at most two over twelfth, eighth, sixth and fourth roots of unity and a golden-ratio pair. What remains open is thus a specific conductor-24, weight-three word problem rather than an unstructured search. PSLQ scans on that residual piece return stable nulls with 45 of 45 positive controls passing. One earlier round of nulls was discarded when its candidate sets proved $\mathbb{Q}$-linearly degenerate. An integer-relation search can satisfy itself on an internal relation of such a set, with zero coefficient on the target, so only nulls on sets checked for that defect are quoted.

S2 Recurrences in the pattern counts

Section 3.2 of the main text mentions a linear recurrence of order fourteen in the constant-site count of the Jukes–Cantor quartet. This section records how such recurrences were found, what their orders are on the three- and four-taxon trees, why they are not used for the values in the paper, and what changes at five taxa.

S2.1 Recurrences along one count

As a function of any one count $y_j \mapsto n$ with the others fixed, Z satisfies a linear recurrence $\sum_{i=0}^{r} c_i(n)\, Z(n+i) = 0$ with $c_i \in \mathbb{Z}[n]$. In the language of A-hypergeometric systems this is a contiguity relation (Gel'fand et al., 1990); such relations in a count direction are also the working tool of the holonomic gradient method for discrete distributions (Takayama et al., 2018). What is new here is the object they act on, an evidence integral over branch lengths, and the measured orders. In words: add one more site of pattern j, and the evidence of the longer alignment is a fixed combination of the evidence values of slightly shorter ones, as Pascal's rule relates neighboring binomial coefficients. The recurrence depends on the model and the tree but not on the data.

We determined the minimal order r and coefficient degree d by evaluating $Z(n)$ modulo 31-bit primes for $n = 0, \ldots, n_{\max}$ and solving the resulting structured nullspace problem for the coefficient combination that makes every evaluation cancel. Working modulo a prime keeps every intermediate number machine-sized; enough primes determine the exact integer coefficients, which the Chinese remainder theorem reassembles. No floating point enters. Operators obtained in this way are reconstructions, not symbolic derivations: each recurrence below is a candidate annihilator that we accept on the finite verifications stated, and none comes with a creative-telescoping certificate (Zeilberger, 1990); the one operator of this supplement that does, and is thereby proved, is the differential operator of Section S1.2. That an annihilating recurrence exists in each count is, however, a theorem of the holonomic systems approach, since the integrand is holonomic. The measured orders are

model	integration dim.	(r, d) range	wall time
three-taxon star	3	$(4, 11)$–$(8, 12)$	$\lesssim 60\,\mathrm{s}$
four-taxon, strict clock	2	$(7, 9)$–$(15, 11)$	$\lesssim 30\,\mathrm{s}$
four-taxon, full	5	$(24, 19)$	$\sim 50\,\mathrm{min}$

Every entry was found at one prime and confirmed at a second. The four-taxon full-model recurrence of order 24 and degree 19 was found at $p_1 = 2147483629$ with 27-row overdetermination and confirmed at $p_2 = 2147483587$ with identical (r, d). For the three-taxon star a recurrence of order 7 and degree 4, minimal in parameter count within its shape family, is given with the programs. The true minimal-order three-taxon operator has order 5 and degree 6, exact over $\mathbb{Q}$ and verified at two primes. Carried from $n = 0$, the order-$(7, 4)$ three-taxon recurrence reproduces the exact $Z(n)$ at $n = 50$ and $n = 70$ to 78 and 79 digits, and reaches $n = 100$, where direct quadrature of the 10^{-66}-peaked integrand fails.

S2.2 The order-fourteen generator

The $(24, 19)$ recurrence is not minimal and its annihilator is not unique. A scan of every shape solvable at series depth 550 finds independent families at $(r, d) = (19, 25)$, $(20, 23)$ and $(26, 18)$, with nullities 3, 2 and 3 that are identical across four primes, so the ideal of valid recurrences has several generators. We lifted all three families to exact integers. Written directly, their coefficients are as large as the determinants of the linear system that produced them, too large to reconstruct entry by entry. Searching the integer solution lattices for their shortest members instead (lattice reduction on the reconstruction of the 156 mod-p nullspaces) gives equivalent recurrences of 2,352, 3,277 and 2,133 bits. Each operator was reconstructed from the values $Z(0), \ldots, Z(550)$. Each annihilates the same 551 values at a further prime withheld from

the reconstruction $(2^{31}-1)$, fails, as it should, when one coefficient is deliberately corrupted, annihilates the first 179 values exactly over $\mathbb{Q}$ (the values themselves reconstructed independently from the mod-p series), and has non-vanishing leading coefficient $c_0(n)$ for $n \le 5000$. The two checks are familiar statistical practice: the withheld prime is data withheld from the fit, and the corrupted coefficient confirms that the test can fail, so that a pass is informative.

The three families are not independent. Their greatest common right divisor in the shift algebra was computed modulo p from the exact operators and lifted by the same route (48 primes suffice). It is a single primitive operator of order 14 and coefficient degree 88 whose integer coefficients, of at most 172 digits, are given in the supplementary file `recurrence_order14.json`. Its 1,335 coefficients exceed the 537 relations that the values to $n = 550$ supply, so it could not have been fitted to them. It was obtained algebraically, as the greatest common right divisor, and then tested: it annihilates $Z(0), \ldots, Z(550)$ at the withheld prime (537 relations), the first 179 values exactly over $\mathbb{Q}$, and $Z(551), \ldots, Z(630)$ at a further prime used nowhere in its construction (80 relations beyond the window). It fails under a corrupted coefficient, and every lifted operator is a right multiple of it at two primes: any sequence obeying the order-14 recurrence obeys all the longer ones. As far as the data reach, this one operator generates the recurrence ideal. Whether order 14 is minimal is open: no lower order exists at any shape solvable at that depth, but coefficient degrees beyond the window are unconstrained.

S2.3 General alignments and five taxa

Seeded with r exact initial values from Lemma 1, a recurrence evaluates $Z(T; \mathbf{y})$ at any N in $O(rN)$ exact rational operations, milliseconds for alignment lengths in the thousands against the 10^7 to 10^8 likelihood evaluations of a stepping-stone run. That speed is available only along the count direction on which the operator was found, here the constant-pattern ray of the order-14 generator. On every mixed-count direction tested (adding k sites of class `xxyy`, $k = 1, 2, 3, 5$) the order-14 operator fails, with 147 of 147 residuals nonzero at two primes. Three ways of acquiring the operator for a general alignment were tried systematically: contiguity matrices, Picard–Fuchs operators, and reduction to a rank-81 compressed system, each validated on three-taxon cases where the known operators are recovered. All three met the same obstacle, that the cost of acquiring the operator grows with the alignment length itself. The evidence is holonomic of modest rank, but at realistic alignment lengths finding the operator has so far cost more than direct exact evaluation. Every real-data value in the paper therefore comes from direct evaluation (Section S3), at the costs of Section 3.3. Closing this gap is an open problem.

At five taxa only the scale changes. The direct route expands a polynomial of degree at most N in each of seven variables, $(N+1)^7$ monomials, about 1.8×10^9 at $N = 20$ and a factor 441 over the quartet. The recurrence route would replace the order-14 quartet generator by the generator of a contiguity ideal in 51 counts, at a holonomic rank we estimate at 50 to 150. That figure is an extrapolation of the measured three- and four-taxon orders, not a computed bound. A scan of the five-taxon constant-pattern ray at windows up to 1001 terms found no recurrence with $(r+1)(d+1) \lesssim 990$, consistent with that estimate. The growth from fifteen pattern classes to 51, and from order 14 to a rank of order 10^2, is the cost of one more internal edge. Just as pruning carries a vector of conditional likelihoods at each node, an exact recursion must carry one value for each independent direction in the larger pattern space. Verification at five taxa is unchanged in kind, exact modular arithmetic against an independent seven-dimensional quadrature; a tropical importance sampler designed for this class of integrals (Borinsky et al., 2023) would give a further independent reference. With two internal edges the degeneration

locus of Section S5.1 becomes a two-parameter object, the first case in which it has geometry of its own.

Two questions about the tempered evidence of Section S1 also remain open. The constant of Eq. (S4) lacks a complete closed form: two of its three pieces are closed, and what remains is the one-dimensional weight-three problem stated there. And it is not known whether the order and singularity structure of the tempered-evidence operator stabilize as the data grow. The two operators in hand grow mildly (order 10, degree 33 at exponents $(1, \frac{1}{2}, \frac{1}{2})$; order 11, degree 37 at $(3, \frac{3}{2}, \frac{3}{2})$), but the limiting parameter is the denominator of βy_k rather than the size of $\mathbf{y}$ (Section S4.1). A statement uniform in $\mathbf{y}$ would turn the deterministic stepping-stone path of Section S4.1 into an algorithm with guarantees.

S3 Evaluation programs and their cross-checks

Algorithm 1 of the main text is the whole of the method. This section describes the three programs that implement it at increasing alignment length and the checks under which every tabulated value was accepted. It then gives the synthetic checks of the Kimura and general time-reversible statements of Section 3.1, and the incomplete graphics-processor run on the 895-site alignment.

S3.1 Three programs

Dense expansion. Direct evaluation of Eq. (6) expands $\prod_k p_k^{y_k}$ over the $(N{+}1)^5$ lattice of per-edge degrees, modulo each prime in turn, and integrates term by term. It is exact but memory-bound, with a practical ceiling near $N \approx 60$ (Algorithm 1, step 2(a)). The twenty-site value of main-text Eq. (12) was computed this way in 14 s.

Transform and stream. A discrete Fourier transform of the coefficient array removes the memory limit. In the match-count form of Eq. (S6) below the array $W_{\mathbf{y}}$ is an N-fold convolution of fifteen fixed 2^5-entry arrays g_k, so its transform over $(\mathbb{Z}/L\mathbb{Z})^5$, for any $L \geq N{+}1$ and computed modulo a prime $p \equiv 1 \pmod{L}$ with a primitive L-th root of unity ω, is the pointwise product $\hat{W}_{\mathbf{y}}(j) = \prod_k g_k(\omega^{j_1}, \ldots, \omega^{j_5})^{y_k}$, and by the discrete Parseval identity the evidence becomes a sum over transform coordinates that can be streamed in $O(L^3)$ memory or less (Algorithm S1). This is a transform of the multi-site exponent lattice, not the Hadamard conjugation of Hendy and Penny (1989) on single-site pattern space. This program was checked against the dense one by exact equality at $N = 20$ and $N = 40$, and against an independent mod-p computation along a one-parameter ray of counts up to $n = 550$. Its cost per prime, L^5 lattice points times a fixed number of modular multiplications with $L = N + 1$, was measured on one core at 0.9, 34 and 343 s for $N = 20$, 40 and 60 and at 1.1×10^4 s for $N = 120$ (from 4 of 121 slices), that is 2 to $4 \times 10^{-7} L^5$ seconds, the coefficient rising slowly as the working arrays outgrow the processor cache. With the primes spread over cores this brings alignments of about 150 sites within reach and leaves full-length alignments of 900 sites beyond these two programs.

Mixed-site reduction. The two pattern classes unchanged by collapsing the cherries of the evaluated topology (`xxxx` and `xxyy` for 12|34) can be separated from the rest. Writing the count vector as $\mathbf{y} = \mathbf{y}_B + \mathbf{y}_A$ with $\mathbf{y}_A$ the m mixed sites, the dependence of Z on the collapsed part is confined to a corner-shaped block of exponent vectors of side $m + 1$. The evidence is then

one exact contraction of three objects: a three-index array for $\mathbf{y}_B$, per-edge ladders of shifted integrals obeying a two-term recurrence, and a five-index array of side $m+1$ for $\mathbf{y}_A$. Everything is computed modulo each prime and reconstructed by the Chinese remainder theorem. The cost is set mainly by m, with the operation count of Eq. (S8) below (main text, Eq. (14)). This program was checked three ways: identical integers with the dense program on the twenty-site case, agreement with the streamed program's independent code, and a full recomputation of eight previously computed tRNA quartets of Table 2. All eight gave identical numerators and denominators on all three topologies in both prime sets, at roughly nine times the aggregate speed of the earlier computations of the same rows. This program produced the lysozyme and 5S values of main-text Table 3 and the timings of main-text Section 3.3; the eight recomputed quartets are rows of Table 2, first computed with the programs above. The twenty-site value is the dense program's, and the quintet of Table 4 was computed by the seven-edge form of the same term-by-term expansion (Section S3.3). The short windows of Sections S6 to S9 have at most eight mixed sites, where the computation is small by any route. The Neandertal values of Section S6, for example, were carried in exact rational arithmetic rather than modulo primes.

S3.2 Algorithm and measured cost

The match-count form. Write the Jukes–Cantor transition probabilities as $P_{ii}(x) = \frac{1}{4}q_1(x)$ and $P_{ij}(x) = \frac{1}{4}q_0(x)$ with $q_1(x) = 1+3x$ and $q_0(x) = 1-x$. In the pruning sum for one site with leaf states s and internal states (a, b), say that edge e *matches* when the states at its two ends agree, and record the indicators $\varepsilon = ([a{=}s_1], [a{=}s_2], [b{=}s_3], [b{=}s_4], [a{=}b])$. Each pattern class k then has a generating polynomial in five indicator variables,

$$g_k(\zeta_1, \ldots, \zeta_5) = \sum_{a,b} \prod_{e=1}^{5} \zeta_e^{\varepsilon_e}, \tag{S5}$$

a $2 \times 2 \times 2 \times 2 \times 2$ array of non-negative integers summing to 16, and collecting, edge by edge, the number m_e of sites at which edge e matches gives

$$\begin{aligned} \prod_k p_k(x)^{y_k} &= 4^{-6N} \sum_{m \in [0,N]^5} W_{\mathbf{y}}(m) \prod_{e=1}^{5} q_1(x_e)^{m_e} q_0(x_e)^{N-m_e}, \\ Z(T;\mathbf{y}) &= 4^{-6N} \sum_{m \in [0,N]^5} W_{\mathbf{y}}(m) \prod_{e=1}^{5} J_N(m_e), \end{aligned} \tag{S6}$$

where $W_{\mathbf{y}}$ is the coefficient array of $\prod_k g_k(\zeta)^{y_k}$, an N-fold convolution of the arrays (S5) whose entries count internal-state assignments and are smaller than 16^N, and

$$J_N(k) = \int_0^1 (1+3x)^k (1-x)^{N-k}\, \mathrm{d}x, \qquad J_N(0) = \frac{1}{N+1}, \qquad J_N(k) = \frac{3k\, J_N(k-1) + 1}{N-k+1} \tag{S7}$$

(integration by parts), rational numbers whose denominators divide $L_{N+1} = \mathrm{lcm}(1, \ldots, N+1)$. Equation (S6) is Eq. (8) of the main text in the basis $\{q_1, q_0\}$ instead of $\{1, x\}$; Algorithm 1 and the two programs below all evaluate it, modulo each prime of Algorithm 1 with J_N replaced by the integers $K(k) = L_{N+1} J_N(k)$, so that the reconstructed integer is $4^{6N} L_{N+1}^5 Z = 16^N D'_N Z$ with D'_N as in Eq. (A1); it is below $4^{5N} L_{N+1}^5$ because $Z < 4^{-N}$, which fixes the number of primes.

Algorithm S1. Transform and stream (any N; memory independent of the $(N{+}1)^5$ array).

Input: counts $\mathbf{y}$ in the frame of T; N. *Output:* $Z(T;\mathbf{y},1)$ as a reduced fraction.

1. Choose a transform length $L \geq N+1$ and primes $p_1,\dots,p_r < 2^{31}$ with $p_i \equiv 1 \pmod{L}$ and $p_1 \cdots p_r > 4^{5N} L_{N+1}^5$; for each p_i find a primitive L-th root of unity ω_i (verify $\omega_i^L = 1$ and $\omega_i^{L/q} \neq 1$ for every prime $q \mid L$). Because every axis degree of $W_{\mathbf{y}}$ is at most $N < L$, the cyclic transform of length L involves no wrap-around; L is otherwise free and is taken large enough that r such primes exist.
2. For each prime p (independently, in parallel): (a) compute $K(0),\dots,K(N)$ by (S7) and $\hat{K}(j) = \sum_{k=0}^{N} K(k)\,\omega^{-jk}$, $j = 0,\dots,L-1$; (b) accumulate

$$R_p = L^{-5} \sum_{j \in [0,L)^5} \Big[\prod_k g_k(\omega^{j_1},\dots,\omega^{j_5})^{y_k} \Big] \prod_{e=1}^{5} \hat{K}(j_e) \bmod p,$$

   streaming over (j_1, j_2) with the (j_3, j_4, j_5) values held as arrays, the powers by binary powering. By the convolution theorem and the discrete Parseval identity, $R_p \equiv 4^{6N} L_{N+1}^5 Z \pmod{p}$.
3. Reconstruct $4^{6N} L_{N+1}^5 Z$ from the R_p by the Chinese remainder theorem as in Algorithm 1, on two disjoint prime sets, and reduce.

Work per prime: L^5 lattice points times (number of classes present + 5) modular multiplications; memory $O(L^3)$ words.

Operation count. Per prime (or, in integer arithmetic, per pass with integers of $O(N \log N)$ bits),

$$\begin{aligned} C(N,m) \;\approx\; & \underbrace{2\,(n_B{+}1)^4}_{\text{build } B} + \underbrace{2\,(m{+}1)^6}_{\text{build } A} + (m{+}1)^5 \\ & + \underbrace{(m{+}1)^2(n_B{+}1)^3 + (m{+}1)^4(n_B{+}1)^2 + (m{+}1)^5(n_B{+}1)}_{\text{stages (a), (b), (c)}}, \qquad n_B = N - m, \end{aligned} \tag{S8}$$

multiply-adds, and the number of primes is $r \approx \log_2(4^{5N} L_{N+1}^5)/31 \approx N/2$ (34 to 40 at the tRNA lengths). Ordering stage (a) as a product over V first lowers its count to $(m{+}1)(n_B{+}1)^3$ and stage (b) to $(m{+}1)^3(n_B{+}1)^2$; building A by the transform of Algorithm S1 would lower its count to about $5(m{+}1)^5 \log_2(m{+}1)$. Neither refinement changes the two regimes: for $m \lesssim 10$ at tRNA length the $(n_B{+}1)^4$ term dominates and the cost hardly depends on m; for larger m the $(m{+}1)^5(n_B{+}1)$ and $(m{+}1)^6$ terms dominate. For the 895-site alignment ($n_B = 694$, $m = 201$) stage (c) alone is $695 \times 202^5 \approx 2.3 \times 10^{14}$ multiply-adds per prime, the figure behind Section S3.5.

Measured cost. Table S1 gives single-core processor time, elapsed time and peak memory of the integer-arithmetic implementation (the released package, one process and one thread per value) on synthetic quartets with $N - m$ constant columns and m columns drawn uniformly from the 240 patterns that are not constant on both cherries, for N varied at $m = 6$ and for m varied at $N = 70$. At $m = 6$ the processor time grows as $N^{4.2}$ between $N = 50$ and $N = 400$ (fitted exponent 4.22 over the whole range and 4.19 over $N = 200$ to 400 alone) and the peak memory as $N^{3.2}$, the regime of the $(n_B{+}1)^4$ term (the three-index array accounts for 80 to 90% of the time; 6.5 core-hours and 20 GB at $N = 400$). At $N = 70$ the processor time is constant to within 10% for $m \leq 7$ (15 to 16 s) and grows as $(m{+}1)^{4.5}$ between $m = 14$ and $m = 30$, where the contraction and, beyond $m \approx 22$, the construction of the five-index array dominate

Algorithm S2. Mixed-site reduction (the program used for the values on real alignments).

Input: counts $\mathbf{y} = \mathbf{y}_B + \mathbf{y}_A$ in the frame of $T = 12|34$, with $\mathbf{y}_B$ the `xxxx` and `xxyy` sites ($n_B = N - m$ of them) and $\mathbf{y}_A$ the m mixed sites. *Output:* $Z(T; \mathbf{y}, 1)$ as a reduced fraction.

1. *Arrays.* For the two classes of $\mathbf{y}_B$ every term of g_k has $\varepsilon_1 = \varepsilon_2$ and $\varepsilon_3 = \varepsilon_4$ (the two leaves of each cherry agree, so they match or mismatch their common node together). Build the three-index array $B[Y, Z, V]$, $0 \le Y, Z, V \le n_B$, of $\prod_{k \in B} g_k^{y_k}$ in the variables $(\zeta_1\zeta_2,\ \zeta_3\zeta_4,\ \zeta_5)$ by n_B successive convolutions with $2 \times 2 \times 2$ kernels, and the five-index array $A[\sigma]$, $\sigma \in [0, m]^5$, of $\prod_{k \notin B} g_k^{y_k}$ by m convolutions with the kernels (S5). Then $W_{\mathbf{y}} = B * A$ and

$$Z = 4^{-6N} \sum_{\sigma \in [0,m]^5} A[\sigma]\, G(\sigma),$$

$$G(\sigma) = \sum_{Y,Z,V} B[Y, Z, V]\, J(Y{+}\sigma_1) J(Y{+}\sigma_2) J(Z{+}\sigma_3) J(Z{+}\sigma_4) J(V{+}\sigma_5).$$

2. *Primes.* As in Algorithm 1, step 1 (no congruence condition).
3. For each prime p, with $K(k) = L_{N+1} J_N(k) \bmod p$ and all arrays reduced modulo p, contract in three stages:
   (a) for each (σ_1, σ_2): $S_1[Z, V] = \sum_Y B[Y, Z, V]\, K(Y{+}\sigma_1) K(Y{+}\sigma_2)$;
   (b) for each (σ_3, σ_4): $S_2[V] = \sum_Z S_1[Z, V]\, K(Z{+}\sigma_3) K(Z{+}\sigma_4)$;
   (c) for each σ_5: $R_p \mathrel{+}= A[\sigma] \sum_V S_2[V]\, K(V{+}\sigma_5)$.

   Each stage is a dense matrix product modulo p with shapes fixed in advance; $R_p \equiv 4^{6N} L_{N+1}^5 Z \pmod p$.
4. Reconstruct and reduce as in Algorithm 1, on two disjoint prime sets.

The released package (Schwartz, 2026) runs steps 1 and 3 once in exact integer arithmetic (entries of B, A below 16^N; K cleared of its common denominator) instead of modulo primes, and returns the same fraction.

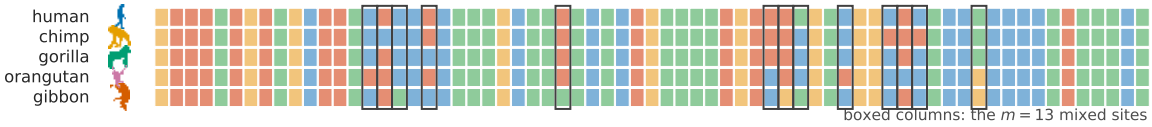

Figure S1: The hominoid mt tRNA-Phe quintet alignment of Section 5 of the main text, all 67 gap-free sites. The 13 boxed columns are the mixed sites, whose count m sets the cost of the exact computation (Section 3.3 of the main text); the remaining 54 are constant.

(17 min and 3.3 GB at $m = 30$). The modular implementation used for the values in this paper distributes the primes over cores; its per-prime times on the eight timed tRNA quartets (Fig. 2 of the main text) range from 0.4 s at $m = 10$ to 530 s at $m = 59$.

S3.3 Cross-checks

Each tabulated exact value computed modulo primes was accepted under the following checks, stated once here for all tables. (i) The integer $D\,Z$ was reconstructed from two disjoint sets of primes, each set large enough on its own by the bound of Lemma 1, and the two reconstructions are identical. For the twelve tRNA quartets the sets hold 34 to 40 primes each, for the 5S alignment 61, for lysozyme 200, and for the quintet 113. (ii) For the quintet, the residue at one further prime withheld from both sets (2147483647) was predicted from the reconstructed fraction and matched. (iii) Every value agrees with floating-point Gauss–Legendre quadrature to the roughly 16 digits that arithmetic carries. The agreement is 15.5 to 16.8 digits on the tRNA quartets, 15.5 to 16.1 on lysozyme and 5S, and 1.4×10^{-13} worst absolute deviation in $\log Z$ on the fifteen quintet topologies. (iv) On the twenty-site case, where higher precision is affordable, nested 11^5-node Gauss–Legendre quadrature at 70 working digits (53 s) agrees with the fraction to 61 digits for 12|34 and to at least 58 digits for the two alternatives. Two programs sharing no method and no code, exact integer arithmetic on one side and high-precision numerical integration on the other, agree digit for digit. An error in either would have to be matched by an identical error in the other to pass. (v) For the twelve tRNA quartets every fraction was re-derived a second time from the stored prime residues, 36 of 36 values identical. The taxon order and alignment coordinates of every case were checked against the source records. (vi) Where a relabeling of the taxa maps the alignment to itself, the program, which has no knowledge of the symmetry, returned identical integers for the exchanged topologies. This occurs in rows 1 and 15 of Table B1 and in the six tied pairs of Table 4. (vii) The six values of Table B1 at prior means 0.1 and 0.01 were instead carried in exact rational arithmetic by the term-by-term formula of Eq. (8) at $\alpha = 15/2$ and 75. The same program returns the integers of row 1 at $\alpha = 1$, the exchanged topologies again give identical integers, and all six values agree with floating-point Gauss–Jacobi quadrature to within 5×10^{-13} log units. Decimal values printed anywhere in the paper are roundings of these exactly known numbers.

S3.4 Other substitution models on synthetic alignments

Lemma 2 of the main text was checked numerically at parameter values chosen to be unfavorable. On the three-taxon star tree under a general time-reversible model with frequencies $\pi = (2, 3, 1, 4)/10$ and exchangeabilities $r = (1, 2, 3, 1, 2, 1)$, the characteristic factor $100\lambda^3 +$

Table S1: Measured cost of exact evaluation by Algorithm S2 in its integer-arithmetic implementation (Schwartz, 2026, version 0.2.3), one core of a shared 2.3-GHz x86-64 server per value (Python 3.12): processor (user CPU) time, elapsed time of the evaluation (larger, because the machine was shared), peak resident memory, the numbers of entries of the three-index array B and the five-index array A, and the digit counts of the reduced fraction. Alignments are synthetic: $N-m$ constant columns and m columns drawn uniformly from the 240 patterns that are constant on neither cherry of 12|34. Upper block, $m = 6$ with N varied; lower block, $N = 70$ with m varied. The three largest alignments of the upper block were run concurrently, each on one core, with a cap of 7.5 hours of elapsed time. Every denominator divides D'_N of Eq. (A1).

N	m	n_B	CPU (s)	elapsed (s)	peak memory (MB)	entries of B	entries of A	digits p/q
50	6	44	4	5	71	45,585	5,582	146/204
60	6	54	8	10	81	83,215	5,395	186/251
75	6	69	20	30	107	171,535	5,557	233/311
100	6	94	72	99	235	428,735	6,243	317/407
125	6	119	169	279	439	864,060	4,958	401/512
150	6	144	387	649	762	1,524,385	6,129	499/624
200	6	194	1308	1823	1805	3,707,535	5,458	678/837
250	6	244	3444	4284	3834	7,353,185	5,145	849/1034
300	6	294	7059	8731	7493	12,836,335	4,844	1008/1236
350	6	344	14525	17035	11917	20,531,985	5,552	1188/1442
400	6	394	23404	26065	20586	30,815,135	4,338	1383/1666
70	2	68	16	23	104	164,289	72	228/286
70	3	67	15	22	102	157,250	275	228/289
70	4	66	15	21	102	150,415	862	229/292
70	5	65	15	22	98	143,781	2,349	223/291
70	6	64	16	18	101	137,345	5,826	215/287
70	7	63	16	21	97	131,104	12,240	217/292
70	8	62	17	25	93	125,055	21,238	212/293
70	9	61	20	29	98	119,195	31,707	210/296
70	10	60	23	31	101	113,521	49,240	209/297
70	11	59	25	39	106	108,030	75,496	208/299
70	12	58	28	44	114	102,719	110,038	206/300
70	14	56	43	58	155	92,625	237,136	207/303
70	16	54	61	88	212	83,215	437,686	202/302
70	18	52	99	148	353	74,465	764,519	198/306
70	20	50	160	228	494	66,351	1,389,792	192/304
70	22	48	239	335	766	58,849	2,094,022	185/304
70	25	45	431	581	1445	48,691	4,114,136	186/310
70	30	40	1032	1224	3308	34,481	10,275,324	181/315

$520\lambda^2 + 861\lambda + 457$ of the rate matrix is irreducible over $\mathbb{Q}$, so all three nonzero eigenvalues are irrational. For the three-site alignment $\mathbf{y} = \{\text{xxx:1}, \text{xxy:1}, \text{xyz:1}\}$ with an exponential branch-length prior the resolvent identity of Eq. (10) gives an explicit fraction with a 76-digit numerator and an 83-digit denominator, in 1.7 s of exact rational arithmetic. A conventional spectral evaluation at 120 and then 220 working digits agrees with it to 119 and 225 digits. On the five-taxon caterpillar tree the same comparison was made twice. Under the Kimura two-parameter model (transition rate 5/4, transversion rate 1/4) with a five-site alignment, term-by-term integration over 4,211,614 monomials gives a 30-digit over 45-digit fraction that the spectral route matches to 58 digits (118 at doubled precision). Under the general time-reversible model at the parameter point above with a four-site alignment, the resolvent route gives a 330-digit over 343-digit fraction that the spectral route matches to 60 digits (118 at doubled precision). Under a gamma branch-length prior of shape $k = \frac{1}{2}$ at $\alpha = 1$ on the three-taxon star, the closed form of Appendix A agrees with direct quadrature to 51 digits. That closed form is an element of $\mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5})$ once the factor $\Gamma(1/2)^3 = \pi^{3/2}$ of the unnormalized prior is divided out. No value on real sequence data was computed under any of these models.

S3.5 A graphics-processor run at full length

Equation (14) places the two 900-site alignments of Table 3, at $m = 201$ and $m = 357$, beyond the programs above on conventional processors. The contraction of the mixed-site program is batched dense integer matrix multiplication modulo a prime, with fixed shapes known in advance, no branching and no floating point. It has independent parallelism within a prime and across primes, topologies and reconstruction sets. Its one remaining bottleneck is memory bandwidth in forming the intermediate array. Graphics processors suit this workload, and because the arithmetic is integer arithmetic modulo a prime, a port gives results identical to the processor version.

A single-processor implementation of this kind was checked against the exact values at reduced scale and started on the 12|34 topology of the 895-site alignment of Brown et al. (1982). It ran at an average of 107 seconds per prime and was stopped after 72 of the 496 residues of one topology and prime set, about a seventh of the set; the run log projected a further 12.6 hours to complete it. No complete exact value at full alignment length exists.

S4 A deterministic stepping-stone path

Section 3.4 of the main text reports a stepping-stone schedule on which every ratio is known without sampling. This section gives the construction in full: the dataset and schedule, the two rational stones, the two stones obtained from differential equations, the wall time, and the limit set by the exponent denominators.

S4.1 Schedule and dataset

Stepping-stone sampling estimates the evidence as a telescoping product over a temperature grid $0 = \beta_0 < \beta_1 < \cdots < \beta_K = 1$,

$$Z \;=\; \prod_{k=0}^{K-1} \frac{Z(\beta_{k+1})}{Z(\beta_k)}, \qquad Z(\beta) = \int_{[0,1]^{|E|}} L(x)^{\beta}\, \pi(x)\, \mathrm{d}x, \tag{S9}$$

with each ratio estimated by importance sampling from the power posterior at β_k (Xie et al., 2011); thermodynamic integration integrates $\mathbb{E}_\beta[\log L]$ over the same grid (Lartillot and Philippe, 2006). Section S1.2 showed that a grid value $Z(\beta_k)$ satisfies a Fuchsian differential equation that can be computed. Here every stone of one complete schedule is evaluated exactly or to a stated precision.

The dataset is chosen so that the grid is tractable: the three-taxon Jukes–Cantor star tree with counts $\mathbf{y}' = \{\text{xxx:}4, \text{xxy:}2, \text{xyz:}2\}$ ($N = 8$ sites), the uniform prior on x_e ($\alpha = 1$), and the uniform $K = 4$ schedule $\beta \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$. The integrand exponents are $\beta \cdot (4, 2, 2)$, so every stone is either integer ($\beta = \frac{1}{2}, 1$, where the evidence is rational by Lemma 1) or half-integer ($\beta = \frac{1}{4}, \frac{3}{4}$, the two-square-root class of Section S1.2). When the exponents are whole numbers the tempered likelihood is still a polynomial in the edge variables and integrates term by term into a fraction. When they are half-integers the integrand contains square roots of the pattern polynomials, the term-by-term route is closed, and the value comes instead from the differential equation it satisfies. Doubling the counts halves the temperature, $\hat{Z}_{\mathbf{y}'}(\beta) = \hat{Z}_{\mathbf{y}}(2\beta)$ with $\mathbf{y} = \{2, 1, 1\}$ the dataset of Section S1.2, so the $\beta = \frac{1}{4}$ stone is the constant of Eq. (S4). The choice of $\mathbf{y}'$ rather than $\mathbf{y}$ is forced. On $\mathbf{y}$ itself the $\beta = \frac{1}{4}$ and $\beta = \frac{3}{4}$ stones carry quarter-integer exponents, for which the mod-p operator search returned no equation at series depths up to 360; probes at exponent denominator $q = 3$ fail in the same way. Operator size grows steeply with q.

S4.2 The four stones

With the rescaled pattern polynomials f_i and $Z_{\mathbf{y}'}(\beta) = 64^{-8\beta}\hat{Z}_{\mathbf{y}'}(\beta)$, term-by-term integration of $\hat{Z}_{\mathbf{y}'}(\beta) = \int_{[0,1]^3} f_0^{4\beta} f_1^{2\beta} f_2^{2\beta}\, \mathrm{d}x$ gives the two rational stones exactly,

$$\hat{Z}_{\mathbf{y}'}(\tfrac{1}{2}) = \frac{17872}{3375}, \qquad \hat{Z}_{\mathbf{y}'}(1) = \frac{50612054}{1157625}, \tag{S10}$$

each reproduced to 33 digits by the independent quadrature of this section (the precision at which its two finest levels agree), and $Z(0) = 1$. The $\beta = \frac{1}{4}$ stone is the 135-digit constant of Eq. (S4). The remaining stone is $\beta = \frac{3}{4}$, with exponents $(3, \frac{3}{2}, \frac{3}{2})$. Deforming all three factors as in Section S1.2, $\hat{Z}(t) = \int \prod_i (1 + t\, g_i)^{w_i \beta}\, \mathrm{d}x$ with $g_i = f_i - 1$ and $w = (4, 2, 2)$, the minimal operator has order 11 and degree 37. It was found from the Taylor coefficients to order 220 computed modulo thirty 25-bit primes (an intermediate recurrence of shape $(8, 21)$ extending each series before the operator search), reconstructed over the integers with a 340-bit rational-reconstruction margin, and verified to annihilate the 220 coefficients at three further primes withheld from the reconstruction, the 40 further coefficients of orders 221 to 260 recomputed from the definition at a prime used nowhere in its construction, and the exact rational coefficients to order 24 computed from the definition (orders 0 to 24, fourteen relations). Unlike the order-10 operator, for which Section S1.2 gives a certificate, it is reconstructed and finitely verified, not proved. Its integer coefficients are given in the supplementary file `tempered_ode_order11.json`. The $\beta = \frac{1}{4}$ equation has order 10 and this one order 11: the exact description grows with the data, so there is no single master formula, only one verified equation per stone. Each equation is then solved as an initial-value problem, carrying the value and its derivatives, known exactly at $t = 0$, along the deformation to $t = 1$. Transporting the exact rational origin jet to $t = 1$ and resumming on the exact exponent ladder gives

$$\hat{Z}_{\mathbf{y}'}(\tfrac{3}{4}) \;=\; 14.592196\ldots, \tag{S11}$$

which agrees with an independent tanh–sinh and Gauss–Legendre quadrature to more than the 40 digits at which that quadrature's two finest levels agree with each other. The transported value is stable to 89 digits between working precisions; we quote it to the 40 digits that the independent route confirms. Wall time for the $\beta = \frac{3}{4}$ stone is 27 minutes on one workstation, 21 to find its operator and 6 to transport it. The $\beta = \frac{1}{4}$ stone reuses the equation of Section S1.2, the rational stones take seconds, and the independent quadrature of the $\beta = \frac{3}{4}$ stone takes a further 19 minutes. The resulting exact $\log Z(\beta_k)$ at the five grid points are the reference against which the stepping-stone and thermodynamic-integration runs of main-text Section 4 (Fig. 4d there) were scored.

The demonstration is deliberately narrow: $N = 8$ sites, three taxa, one dataset, and a grid chosen to avoid exponent denominators $q \geq 3$. The operator-search cost above is a measured limit. A general-purpose deterministic replacement for stepping-stone sampling would need either operator methods that survive larger q or schedules restricted to half-integer exponents as here. In words, the difficulty of a temperature is set by the arithmetic of β times the counts, not by how much data there is. Half-integer powers, which are square roots, are within reach, and thirds and beyond are not yet.

S4.3 Monte Carlo error and alignment length

Main-text Section 3.3 states that a sampler's Monte Carlo error at a fixed number of generations depends only weakly on the alignment length N. Two measurements support this, and a third measures how the discretization bias of thermodynamic integration depends on N. Multiplying every pattern count by c replaces L by L^c, so the power posterior of the multiplied alignment at β is that of the original at $c\beta$ and $Z_{c\mathbf{y}}(\beta) = Z_{\mathbf{y}}(c\beta)$ exactly (the identity used for the $\beta = \frac{1}{4}$ stone above); a stepping-stone run on the multiplied alignment is a run on the original with the schedule stretched to $[0, c]$.

First, the variance of $\log L$ under the power posterior, which sets the variance of each stepping-stone ratio and of each thermodynamic-integration ordinate, was measured from the MrBayes samples at each stone (Table S2). For $\beta \geq 0.1$ it is the same, within 15 %, on the tRNA-Gln quartet ($N = 68$) and on the hominid tRNA-Phe quartet with its counts multiplied by one, two and four ($N = 69$, 138, 276), and close to $d/(2\beta^2)$ with $d = 5$, the value for a likelihood that is locally Gaussian in the five branch lengths. Only at the stones nearest the prior does it grow with N, and there as N^2 ($1805 : 7293 : 28903 = 1 : 4.04 : 16.0$), because under the prior $\log L$ is N times a site average whose spread does not depend on N.

Second, the dependence on N was tested directly, by stepping-stone runs in MrBayes 3.2.7a on the three multiplied alignments, topology 12|34, at three settings (50 stones of 50,000 or 400,000 generations, 200 stones of 50,500), four seeds of two independent runs each, scored against the exact value at $c = 1$ and against tensor Gauss–Legendre values at $c = 2$ and 4 ($\log Z = -311.4641853$ and -608.3256200, stable to 10^{-9} between 60 and 70 nodes per dimension; the same quadrature reproduces the exact $c = 1$ value to ten decimals). The runs were made with the node-slider proposal disabled (Section S4.4); with the default proposals every cell carries the additional offset of that section, -0.022 to -0.043 log units, at all three N. Table S3 gives bias, standard deviation and root-mean-square error. At each setting the standard deviation is the same at $N = 69$, 138 and 276 within the sampling uncertainty of an eight-run estimate (about a quarter), whereas an error growing as $\sqrt{N}$ would double. The bias is within error of zero at the two larger budgets, and the wall time per run (11 to 14 s, 92 to 124 s and 42 to 51 s at the three settings, on a shared machine) does not depend on N, because the multiplied

Table S2: Variance of $\log L$ under the power posterior at selected stones of the Beta$(0.4, 1)$, $K = 50$ schedule, from 3,600 to 4,800 MrBayes samples per stone (400,000 generations per stone, first quarter discarded, runs pooled). Last row: the large-sample value $d/(2\beta^2)$ for $d = 5$ branch lengths.

β	0.951	0.572	0.279	0.101	0.0179	0.0032	0
tRNA-Gln, $N = 68$	3.2	8.6	35.7	182	568	604	614
tRNA-Phe $\times 1$, $N = 69$	3.6	10.1	45.5	316	1725	1852	1805
tRNA-Phe $\times 2$, $N = 138$	3.5	9.8	43.9	340	5243	7280	7293
tRNA-Phe $\times 4$, $N = 276$	3.4	9.5	39.5	337	8989	28691	28903
$d/(2\beta^2)$	2.8	7.6	32.2	244	7813	2.5×10^5	∞

alignments compress to the same eleven site patterns. Between settings the standard deviation falls at least as fast as the inverse square root of the number of generations.

Third, the discretization bias of thermodynamic integration was measured directly. Runs of RevBayes 1.4.0 were made on the same multiplied alignments and on an eightfold one ($c = 8$, $N = 552$; reference $\log Z = -1199.9093092$ by the same quadrature, stable to 10^{-6} between 50 and 70 nodes per dimension) with the settings of Section 4.3 ($K = 50$ powers $\beta_i = \{i/(K-1)\}^{5/2}$, 10,000 generations per power), at $K = 200$, and at $K = 50$ with 80,000 generations per power, eight seeds per cell. Path-sampling and stepping-stone estimates come from the same samples, so their paired difference estimates the bias free of most of the Monte Carlo noise. That bias was also computed deterministically, as the trapezoid sum of $\mathrm{E}_\beta[\log L]$ over the schedule minus $\log Z$, with $\mathrm{E}_\beta[\log L]$ at every power by the tensor quadrature (Table S4). The two agree within 0.006 log units in every cell. At $K = 50$ the bias is -0.018, -0.033, -0.059 and -0.104 log units at $N = 69$, 138, 276 and 552: it grows by a factor of about 1.8 per doubling of N, less than proportionally. Split by power, the intervals with $\beta \geq 0.1$ contribute -0.010 at every N, as expected when the variance of $\log L$ there does not depend on N (Table S2). The intervals below $\beta = 0.1$ contribute -0.008, -0.023, -0.048 and -0.094, nearly in proportion to N, with the largest single-interval error at $\beta = 0.071$, 0.036, 0.019 and 0.011, that is, at fixed $c\beta$, where the power posterior crosses over from the prior. At $K = 200$ every deterministic bias is smaller by the factor $\{(200-1)/(50-1)\}^2 = 16.5$ at all four N, the K^{-2} law, and the measured paired differences follow it. The bias is the same at 10,000 and 80,000 generations per power, and the stepping-stone estimates from the same runs are within two standard errors of the reference in ten of the twelve cells (the other two, -0.0056 ± 0.0018 at $c = 1$, $K = 50$ and $+0.0073 \pm 0.0037$ at $c = 1$, $K = 200$, are of opposite sign at the same N), with no trend in N.

S4.4 Localizing the MrBayes offset

Main-text Section 4.3 reports a residual offset of about -0.016 log units in the MrBayes 3.2.7a stepping-stone estimates on the tRNA-Gln topologies, and one of the same sign on the tRNA-Phe alignments, that persists when the number of stones (50 to 200) or the chain length (one to eight times the baseline) is increased. All runs below used the MrBayes binary of the main text (Ubuntu package 3.2.7a-7build2); standard errors are over independent estimates.

The offset is unchanged, within its standard error, by reversing the direction of the stepping-stone path, doubling the initial or the within-stone burn-in, and double- in place of single-precision likelihoods (-0.016 to -0.029 log units in each variant, standard errors 0.002 to 0.005), and it does not shrink with alignment length (-0.036, standard error 0.004, on a 390-site JC69

Table S3: Stepping-stone error against alignment length. Hominid tRNA-Phe quartet, topology 12|34, JC69, $\alpha = 1$, pattern counts multiplied by c; MrBayes 3.2.7a with the node-slider proposal disabled, K stones at Beta$(0.4, 1)$ spacing, eight estimates per cell (four seeds, two independent runs each). Bias is the mean estimate minus the reference $\log Z$ (exact at $c = 1$, quadrature at $c = 2, 4$), with its standard error; SD is the standard deviation across the eight estimates; RMSE the root-mean-square deviation from the reference; wall, mean seconds per run of two chains on one core.

c	N	K	gen./stone	bias	s.e.	SD	RMSE	wall (s)
1	69	50	5×10^4	-0.006	0.010	0.027	0.026	12
2	138	50	5×10^4	-0.019	0.014	0.041	0.043	14
4	276	50	5×10^4	-0.016	0.008	0.024	0.027	14
1	69	50	4×10^5	-0.0030	0.0024	0.0066	0.0069	96
2	138	50	4×10^5	$+0.0003$	0.0020	0.0056	0.0052	92
4	276	50	4×10^5	-0.0013	0.0020	0.0057	0.0055	102
1	69	200	5.05×10^4	$+0.0041$	0.0058	0.0165	0.0159	47
2	138	200	5.05×10^4	-0.0008	0.0047	0.0132	0.0124	49
4	276	200	5.05×10^4	$+0.0040$	0.0063	0.0179	0.0172	42

Table S4: Thermodynamic-integration (path-sampling) bias against alignment length at fixed K. Hominid tRNA-Phe quartet, topology 12|34, JC69, $\alpha = 1$, pattern counts multiplied by c; RevBayes 1.4.0, K powers $\beta_i = \{i/(K-1)\}^{5/2}$, eight seeds per cell. PS and SS: mean path-sampling and stepping-stone estimate minus the reference $\log Z$ (exact at $c = 1$, quadrature at $c = 2, 4, 8$), with standard errors; PS$-$SS: mean paired difference; TI_∞: the deterministic discretization bias (trapezoid sum of $\mathrm{E}_\beta[\log L]$ by quadrature minus $\log Z$), in total and from the powers $\beta \geq 0.1$ alone. Log units.

c	N	K	gen./power	PS (s.e.)	SS (s.e.)	PS$-$SS (s.e.)	TI_∞	TI_∞, $\beta \geq 0.1$
1	69	50	10^4	$-0.0158\,(0.0183)$	$+0.0029\,(0.0171)$	$-0.0187\,(0.0019)$	-0.0182	-0.0100
2	138	50	10^4	$-0.0218\,(0.0123)$	$+0.0139\,(0.0109)$	$-0.0357\,(0.0029)$	-0.0330	-0.0105
4	276	50	10^4	$-0.0580\,(0.0158)$	$-0.0043\,(0.0147)$	$-0.0537\,(0.0024)$	-0.0588	-0.0103
8	552	50	10^4	$-0.1114\,(0.0362)$	$-0.0089\,(0.0318)$	$-0.1025\,(0.0071)$	-0.1035	-0.0100
1	69	50	8×10^4	$-0.0241\,(0.0018)$	$-0.0056\,(0.0018)$	$-0.0184\,(0.0006)$	-0.0182	-0.0100
2	138	50	8×10^4	$-0.0351\,(0.0061)$	$+0.0002\,(0.0056)$	$-0.0353\,(0.0007)$	-0.0330	-0.0105
4	276	50	8×10^4	$-0.0509\,(0.0087)$	$+0.0058\,(0.0085)$	$-0.0567\,(0.0016)$	-0.0588	-0.0103
8	552	50	8×10^4	$-0.0948\,(0.0092)$	$+0.0068\,(0.0090)$	$-0.1017\,(0.0026)$	-0.1035	-0.0100
1	69	200	10^4	$+0.0072\,(0.0037)$	$+0.0073\,(0.0037)$	$-0.0002\,(0.0002)$	-0.0011	-0.0006
2	138	200	10^4	$-0.0110\,(0.0055)$	$-0.0093\,(0.0057)$	$-0.0017\,(0.0004)$	-0.0020	-0.0006
4	276	200	10^4	$-0.0138\,(0.0123)$	$-0.0114\,(0.0120)$	$-0.0024\,(0.0009)$	-0.0036	-0.0006
8	552	200	10^4	$-0.0181\,(0.0097)$	$-0.0137\,(0.0093)$	$-0.0043\,(0.0007)$	-0.0063	-0.0006

alignment). At the branch-length vectors MrBayes samples, its printed log-likelihood agrees with an independent pruning computation to 10^{-5} log units and its log prior with the stated exponential density to 10^{-9}, and a run with a constant likelihood returns a log marginal likelihood of exactly zero, so the likelihood, the prior bookkeeping and the stepping-stone arithmetic are not the source. Compared stone by stone with per-stone ratios computed by quadrature, the MrBayes contributions fall short by 3 to 10×10^{-4} log units at every stone with β above 0.01 and by almost nothing below it, as expected if the chain samples slightly longer trees at every power of the likelihood.

On a fixed topology MrBayes 3.2 updates branch lengths with three proposals, `Multiplier(V)` (66.7 per cent of proposals), `Nodeslider(V)` (23.3 per cent) and `TLMultiplier(V)` (10 per cent). With `propset Nodeslider(V)$prob=0`, issued after any `mcmcp` command that changes `nruns`, the offset disappears. Over eight of the exact problems of this and the companion paper, with 72 estimates per setting, the mean deviation from the exact values is -0.0260 log units (standard error 0.0027) with the default proposals and -0.0026 (0.0019) without the node slider, with the same sign in every problem; an earlier set of 70 estimates per setting gave -0.0268 (0.0014) and -0.0002 (0.0013). On the three tRNA-Gln topologies separately the default offsets are -0.0164, -0.0170 and -0.0129 (0.0032 to 0.0035), equal within error, so the offset largely cancels in Bayes factors between topologies. RevBayes, whose branch-length proposals are per-branch scale moves, shows no such offset (main-text Fig. 4).

The node slider is `Move_NodeSlider` in `src/proposal.c` of the MrBayes source (lines 7982 and 7989 at tags v3.2.7 and v3.2.7a; lines 8420 and 8427 at tag v3.2.8 and in the development branch as of 18 September 2026). It multiplies the sum M of two adjacent branch lengths by $c = e^{\lambda(u-1/2)}$, redraws one of the two lengths uniformly on the admissible window of width W, sets the other to the remainder, and assigns the log proposal ratio $\ln(W'/W) + 2\ln(M'/M)$. Because the second length is redrawn rather than rescaled, the map from (M', x'_q) to the new pair has unit Jacobian and the Hastings ratio is $M'W'/(MW)$, that is $\ln(W'/W) + \ln(M'/M)$, which is the expression in MrBayes versions 3.1.2 to 3.2.3; the second factor of M'/M entered with release 3.2.4. Used alone, the kernel as coded satisfies detailed balance with respect to the target density multiplied by M. A direct re-implementation of the routine, applied to two independent exponential branch lengths of rate 4/3, samples their sum from a Gamma(3) distribution instead of Gamma(2) (mean 2.247, standard error 0.005, against 2.25; Kolmogorov–Smirnov distance to Gamma(3) below 0.003) and returns Gamma(2) when the factor 2 is replaced by 1 (mean 1.499, standard error 0.004, against 1.5). MrBayes itself shows the same behaviour when run without data. With four taxa, the topology fixed, `prset brlenspr=unconstrained:exponential(1.0)`, the other two branch-length proposals disabled and `mcmc data=no`, version 3.2.7a samples a mean tree length of 7.03 (standard error 0.01; eight chains of 8×10^6 generations) against the prior's 5, every branch 1.40 to 1.41 instead of 1. At rate 10 the sampled mean is 0.703 (0.001) against 0.5, and under the default `unconstrained:gammadir(1.0,0.1,1.0,1.0)` it is 30.3 (0.1) against 10. With the node slider disabled the prior is recovered (ratio of sampled to prior mean 1.0004 to 1.0005), and in the default mixture of the three proposals the sampled prior tree length is 0.2 to 0.5 per cent too long. On our alignments (4 to 5 taxa, 67 to 390 sites) the posterior mean tree length with the move on exceeds that with it off by 0.05 to 0.11 per cent, about 0.005 posterior standard deviations. The effect is far below any model-choice threshold and does not bear on trees or branch lengths inferred with MrBayes in ordinary use.

S5 Where the quartet integrand degenerates

The exact ties of the main text (Eq. (17) and Table 4) arise from symmetries of the pattern counts. A second source of degeneracy lies in the branch lengths. This section determines, for the Jukes–Cantor quartet, every value of the internal-edge variable at which the constant-site pattern hypersurface, which is a factor of the evidence integrand of every alignment, becomes singular, as an explicit algebraic variety. Only two of those values lie in the range of real branch lengths, a zero and an infinite internal branch, and they mean different things for inference: at the first the three topologies prescribe the same pattern distribution, so that the split is not identifiable, whereas at the second the split remains identifiable and what is lost is the separate identifiability of the pendant lengths within each cherry. The section then counts the critical points of the integrand at generic exponents and at the twenty-site data of the main text.

S5.1 The degeneration locus

The simplest members of the set are familiar: an internal branch of length zero, where the quartet is a star tree, or one so long that the sites are randomized across it, where the two cherries evolve independently. The computation below lists every value of the internal-edge variable at which the constant-site hypersurface or one of its face truncations is singular, including values not visible by inspection. The set is a property of the data configuration and the model, not of any algorithm, and the tool that computes it is the Landau analysis of Feynman integrals.

Consider first integration over closed complex contours with generic coefficients. There the singularities of an integral of the form of Eq. (6) in its parameters are confined to the principal A-determinant E_A of the associated GKZ system, the Landau variety (Fevola et al., 2024a,b). Two hypotheses of that confinement theorem fail here, in instructive ways. The evidence integral runs over the unit box, whose $x_e = 1$ faces carry no twist (the prior factor $x_e^{\alpha-1}$ is inert there), so degenerations at those endpoint faces can escape E_A. And at the Jukes–Cantor point the specialization of the generic E_A vanishes identically, because the coefficients lie on the A-discriminant (the nodal lines below). The locus displayed next is therefore the output of the principal-A-determinant computation (Fevola et al., 2024b) run on the specialized polynomial, a computation that stands on its own, and not an instance of an a priori confinement theorem.

Two computations determine the object. First, the principal A-determinant of the constant-site pattern polynomial along the internal-edge variable x_5 factors into seven simple components,

$$E_A \;\propto\; x_5\,(x_5-1)\,(x_5+\tfrac{1}{3})\,(x_5+\tfrac{3}{4})\,(x_5+\tfrac{1}{2})\,(x_5-\tfrac{3}{2})\,(x_5+3), \tag{S12}$$

and the face stratification of the Newton polytope labels each factor by origin. The polytope is the solid whose corners are the occurring exponent vectors, and its faces index the integrand's limiting regimes. The bulk discriminant contributes $x_5 \in \{1, -\frac{1}{3}\}$: $x_5 = 1$ is a zero-length internal branch, and $-\frac{1}{3}$, like the cherry values, lies outside the range of real branch lengths. The two-edge cherry faces contribute $x_5 \in \{\frac{3}{2}, -\frac{1}{2}\}$, and the vertices contribute the remainder. Second, the full five-variable singular locus of the same hypersurface decomposes into eight irreducible curves, four of which are nodal lines present at every value of the internal edge. The Jukes–Cantor quartet coefficient vector thus lies on the A-discriminant of its Newton polytope, and the constant-pattern hypersurface is generically singular. At four taxa, that is, the degeneracy is not confined to isolated branch lengths but persists along whole curves in branch-length space, whatever the internal branch. The three-taxon analog is smooth (its Jacobian ideal is

the unit ideal), so the phenomenon first appears at four taxa, with the internal edge. Every component of the locus is governed by the same two values $\{1, -\frac{1}{3}\}$ and the cherry constraint $3x_ix_j + 1 = 0$.

The map back to data reads off the factorization directly. On the scale $x_e = e^{-\mu t_e} \in [0, 1]$ of Section 2 of the main text, exactly two of the seven components meet the unit cube $[0, 1]^5$. They are $x_5 = 1$ and $x_5 = 0$. At $x_5 = 1$, an internal branch of length zero, the three topologies give the same pattern distribution for the same pendant lengths (the star tree), so that data generated there carry no signal separating them and the log Bayes factors between topologies do not grow with the number of sites. At $x_5 = 0$, an internal branch so long that the sites are randomized across it, the pattern distribution of $12|34$ factorizes into two independent cherries, $p_s = \frac{1}{16} P_{s_1 s_2}(x_1 x_2)\, P_{s_3 s_4}(x_3 x_4)$ in the notation of main-text Eq. (4): leaves $1, 2$ agree with probability $(1 + 3x_1x_2)/4$, leaves $3, 4$ with probability $(1 + 3x_3x_4)/4$, and every cross pair with probability $1/4$ exactly. No other topology produces this distribution at real branch lengths unless one cherry is itself saturated ($x_1x_2 = 0$ or $x_3x_4 = 0$), because under $13|24$ a cross-pair agreement of $1/4$ forces a pendant variable to vanish and with it one of the within-cherry correlations. The split therefore remains identifiable at $x_5 = 0$, and strongly so; what degenerates there is the parametrization, whose Jacobian drops from rank five to rank three, so that only the products x_1x_2 and x_3x_4 are identified. The other five roots, the cherry pair $\{\frac{3}{2}, -\frac{1}{2}\}$ among them, lie outside $[0, 1]$, and the cherry nodal lines never enter the cube, since $3x_ix_j + 1 \geq 1$ for $x_i, x_j \in [0, 1]$. They shape the analytic structure of the integral as a function of the internal edge but correspond to no realizable data. Among degenerations in the internal-edge variable, then, these two are the only reachable ones. They do not exhaust the ways in which quartet data can fail to separate topologies: a saturated pendant edge ($x_i = 0$) makes leaf i independent of the rest and leaves a three-taxon tree that all three topologies realize, and the label symmetries of Section 4 of the main text tie topologies exactly at any branch lengths. For generic coefficients on the same Newton polytope the complement has signed Euler characteristic 16, the invariant that fixes the holonomic rank (Huh, 2013). The Jukes–Cantor point is not generic, so 16 is not the rank of the system at hand.

S5.2 Critical-point counts

The two counts below answer different questions and are not comparable. At generic exponents a monodromy computation (numerical homotopy continuation) at one generic parameter point found 25,923 nondegenerate critical points of the full quartet master function, 25,918 of which were tracked to a rational exponent draw. Neither set was certified and no completeness test was applied, so this is an uncertified count that we read as indicating a generic holonomic rank of order 10^4 rather than 16. At the twenty-site data of main-text Eq. (12), regulated by prior exponents $\varepsilon = 10^{-3}$ and 5×10^{-4}, 1,603 distinct nondegenerate critical points are certified by Krawczyk's interval test at exact rational parameters, the same number at both regulators; this is a rigorous lower bound. We found no others: the solutions were obtained by parameter homotopy from the solution set of the same system at generic exponents, which was stable under monodromy at two independent parameter draws (2,506 solutions each), but that stability is evidence of completeness rather than proof, and we have not matched the count to an independent root bound. Of these 413 are real and exactly one lies in the unit cube. The Jukes–Cantor data point is thus far below the generic complexity of its own integral class. The drop on the discriminant is quantified at four taxa on a sub-arrangement, with 831 critical points at generically perturbed coefficients against 122 at the Jukes–Cantor coefficients; both of these

counts are complete, each being the number of solutions of the saturated critical ideal read from a Gröbner basis modulo two primes, and the second agrees with a certified homotopy count. For an analyst each count has one consequence. The generic count says how complex an evidence integral of this shape can be. The count at the data says that at real Jukes–Cantor data the integrand has a single interior peak, so the complicated critical-point geometry lies outside the region of integration.

We are not aware of a prior computation of this locus. Its components are the values of the internal edge at which the analytic structure of the quartet integral changes. One of its two real points, the star tree, is also a point of topological non-identifiability, which places it beside the identifiability varieties of Allman and Rhodes (2006) for maximum likelihood and the real-asymptotic analyses of singular marginal likelihoods for star models (Rusakov and Geiger, 2005).

S6 The 1997 Neandertal alignment

Section 4 of the main text reports that, on the alignment of Krings et al. (1997), two modern human sequences pair against two Neandertal sequences by 25.1717 log units under JC69, with the two alternative pairings exactly tied. This section gives the data recovery, the quartets, the controls, and a check of model adequacy. Every value here is a four-taxon Jukes–Cantor evidence under the prior of the main text ($\alpha = 1$), computed in exact rational arithmetic. The section restates the 1997 separation under one model and prior and makes no new claim about hominin history.

S6.1 Data

Krings et al. (1997) published 379 bases of mitochondrial hypervariable-region I sequence from the Neandertal type specimen (Feldhofer 1). The 1997 article had no machine-readable supplement, and its Figure 2 is the published alignment: the reference row and 49 clone rows over 62 columns (the printed window plus one insertion column). It was transcribed from the printed figure and checked cell by cell against the reference sequence and the GenBank Neandertal record. The analysis window is the published one, 61 sites once the insertion column is dropped (JC69 has no insertion–deletion process). Outside the heteroplasmic C-run that the paper also excludes, the Neandertal sequence differs from the human reference (Anderson et al., 1981) at 27 positions, 24 transitions, two transversions and one insertion, exactly the paper's stated count. The paper's own modern-human contaminant clones supply human lineages beside the reference, so every sequence in the quartets below is one the 1997 paper itself printed.

Five quartets were evaluated on the 61-site window: two decision quartets, two controls and one with a planted substitution. Each decision quartet seats the reference sequence, one contaminant human clone, the reported Neandertal sequence and one independent Neandertal amplification clone; the two decision quartets use different human and Neandertal clones (A10.18 with A17.1; B11.11 with A10.5). On every topology of every quartet the number of mixed sites is at most eight, small enough that the values were carried in exact rational arithmetic rather than modulo primes (Section S3). Two independent runs returned identical fractions. Every value satisfies the denominator bound of Lemma 1, $\mathrm{den}(Z) \mid 256^{61}\ \mathrm{lcm}(1,\dots,62)^5$, and permuting the leaves within a cherry returns the identical fraction. The chimpanzee outgroup alignment and the 2,051-sequence modern-human comparison panel of the 1997 analysis resided in a database published only the following year. The rooted form of the 1997 statement therefore cannot be

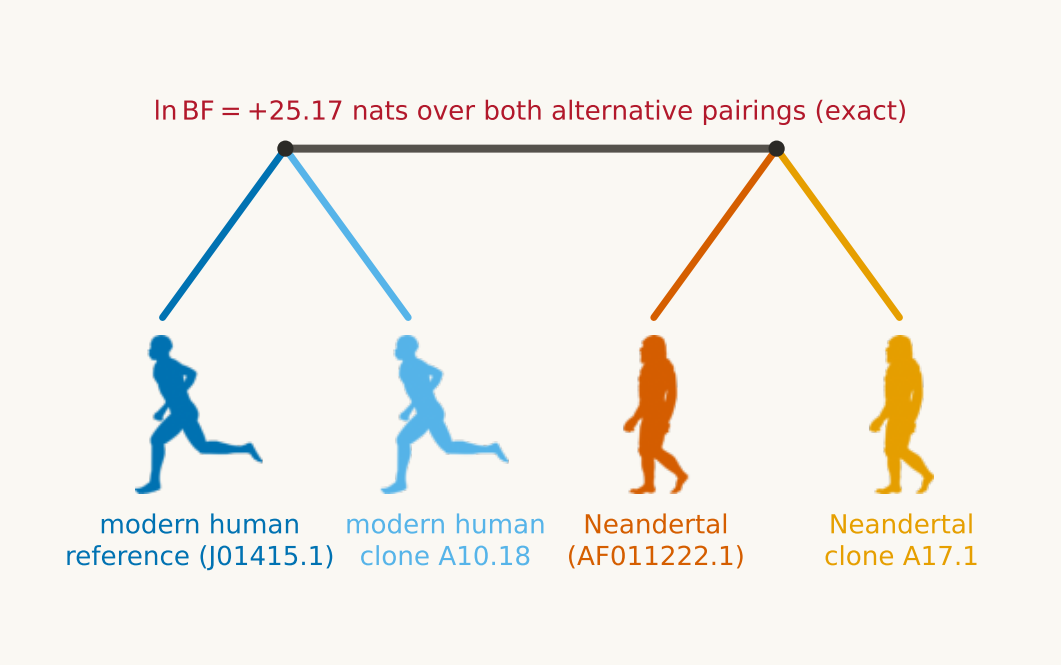

Figure S2: On the 1997 alignment the two modern human sequences (the reference and a contaminant clone from the paper's Figure 2) pair together and the two Neandertal sequences (the reported sequence and an independent amplification clone) pair together, by $\log \mathrm{BF} = +25.17$ log units over each alternative pairing; the two alternatives are exactly tied. JC69, $\alpha = 1$, 61 sites. The tree is drawn unrooted because the exact statement concerns pairings only; the outgroup alignment of the 1997 analysis was not published with the paper.

recomputed from the paper as published, and only the unrooted four-sequence separation is treated here.

S6.2 Result and controls

On both decision quartets the humans | Neandertals pairing is preferred with $\log \mathrm{BF} = +25.171746$ against each of the two alternative pairings (Fig. S2). The Bayes factor is an exact integer ratio, a 199-digit numerator over a 188-digit denominator, about 8.5×10^{10}. The two decision quartets give the identical fraction, as they must: the independent clones resolve to identical sequences on this window, so the two quartets have the same site-pattern counts. For the same reason the two losing pairings are exactly tied. The Neandertal sequence and the Neandertal clone are identical over these 61 sites, so exchanging them maps the alignment to itself while exchanging the two mixed pairings. The argument of main-text Eq. (17) then forces equal marginal likelihoods before any number is computed, and the computed fractions are equal.

Three controls accompany the result. The all-Neandertal quartet (three Neandertal-class sequences and the reference) is an exact three-way tie, every ratio exactly 1/1, again forced by the identity of the Neandertal-class sequences and confirmed by the computed values. Planting one substitution in the human clone (A to G at position 16,230) changes every computed value and lowers the margin from 25.17 to 20.34 log units, in the direction it should. A human-side control quartet seats the reference and one human clone against a second human clone and the

Neandertal. It prefers that pairing by only +2.94 log units against the closer alternative. That preference traces to a single state (C to T at 16,223) that the second human clone shares with the Neandertal. The two losing alternatives of this control differ by 0.035 log units, a difference the exact values resolve.

S6.3 Model adequacy

The exact values also measure how far JC69 falls short on these data. On each of the five quartets the JC69 evidence lies 0.65 to 0.68 log units per site below the saturated multinomial model (Goldman, 1993), an upper bound on what any independent-sites model could gain. A posterior-predictive comparison of the site-pattern spectrum gives a discrepancy of $D \approx 0.045$ to 0.065 log units per site, with an excess of constant sites ($53/61 = 0.869$ observed against 0.805 predicted). This is the pattern-level signature of rate variation across the sites of this hypervariable region. Analyses of the same quartets under the Kimura two-parameter and general Markov models require sampling, are not exact, and are not part of this supplement (Schwartz, 2026).

S7 San Cristóbal tortoise protocol

Section 6 of the main text tests the published placement of the extinct San Cristóbal giant tortoise with a sum of twelve exact log Bayes factors under a design fixed in advance. This section gives the design, the data, the planted-damage null distribution, the result, a calibration of the decision rule under both placements, and the checks that qualify it. Every evidence value is a four-taxon Jukes–Cantor evidence computed exactly as in Section S3; the damage replicates and the summed statistic are built from such values and involve no sampling; the calibration of Section S7.4 uses simulated alignments scored in floating point at a validated accuracy of 1.5×10^{-6} log units.

S7.1 Question and design

Jensen et al. (2022) sequenced tortoise bones collected in 1906 in a cave on San Cristóbal (Galápagos). From one mitochondrial control-region locus of about 700 nucleotides they placed the extinct cave lineage sister to the species pair *Chelonoidis hoodensis* and *C. abingdonii*, distinct from the island's living tortoises, with posterior probability 1.0. The published probability came from a clock-model Bayesian tree search across 93 haplotypes, a different computation from the one made here. The name *chathamensis* is attached to a cave-bone holotype, so the placement decides which population carries the name. The current checklist of the Turtle Taxonomy Working Group (2025) takes no taxonomic act and makes the question conditional on nuclear data, none of which has been recovered from the cave bones. This mitochondrial locus is therefore the only placement evidence available for these bones. The question asked here is not which placement is right, but whether the locus supports the published placement once the null hypothesis includes the damage that ancient DNA accumulates.

The design was written and fixed before any evidence value on the real data existed. It reads five sequences: the holotype cave-bone fragment, a modern eastern San Cristóbal tortoise, the two sister-pair species, and an outgroup tortoise. By exact tree induction the five-taxon statement is the conjunction of two quartet splits, so two decision quartets are read, each on six prespecified 118-column tiles of the 708-column alignment. The declared statistic S is the sum of the twelve exact log Bayes factors of the published-placement split over the alternative split on

which the cave lineage joins the modern islander. Support for the published placement required $S \geq 3.0$ and, in addition, that the two quartet partial sums agree in sign; S is used only for this calibrated decision and is never quoted as a combined evidence. The null distribution plants C→T and G→A changes, the substitution classes that post-mortem deamination produces, on the modern islander's sequence masked to the cave bone's coverage. The planting rate is one, two and five times a base rate of 0.01 per site, taken from the upper end of the published damage range for museum-age material (Sawyer et al., 2012; Hofreiter et al., 2001). The top rung reaches the scale of the real ancient–modern divergence (18 differing sites). Fifty replicates per rung were planted and evaluated before any real-data cell was computed; a rung counts as clean only if none of its fifty replicates reaches the real statistic. Two further rules were fixed in advance. A zero-change replicate, in which the masked modern sequence stands in for the cave bone with no change planted, was evaluated first as a check on the construction of the cells. Its first two rows are identical, so the two site-pattern classes that alone favor separating the cave lineage from the modern islander ($xyxy$ and $xyyx$, in the leaf order cave bone, modern islander, sister species, outgroup) must have count zero in every one of the twelve decision cells, and the program's own count record was required to show this. The same identity forces a tie between two of the three marginal likelihoods, though not the tie that would make S vanish. Exchanging the labels of the two identical rows leaves every site-pattern count unchanged, fixes 12|34 and carries 13|24 into 14|23, so that $Z(13|24) = Z(14|23)$ exactly by the argument of main-text Eq. (17), whereas $Z(12|34)$, the split that joins the two identical rows, is larger whenever a column groups them against the other two sequences. On the zero-change replicate S is therefore not zero but -54.30, the value the statistic takes when the cave-bone sequence is a copy of the modern islander's over the cave bone's coverage; the planted-damage replicates depart from this value. If more than two of the twelve real cells could not be computed the test would be reported as not evaluable. The design thus allowed three outcomes. The first is support ($S \geq 3.0$ with both quartet sums of the same sign). The second is a shortfall that planted damage reproduces ($S < 3.0$ and at least one damage replicate reaches S). The third is a shortfall that damage does not reproduce.

S7.2 Data

Every input is a public GenBank or RefSeq record: six cave-bone fragments, four modern San Cristóbal records (two of them mitochondrial genomes), the two sister-pair reference mitochondrial genomes, and an outgroup. They were aligned once; the alignment used is the 708-column union footprint of the six cave-bone fragments, thirteen sequences by 708 columns. Among the five sequences the design reads there are 14 parsimony-informative sites. No raw sequencing reads exist for these bones; the published sequences are Sanger consensus from amplified fragments. The damage null is therefore a substitute constructed at and above published damage rates rather than measured from reads, and this qualification applies to every number below.

S7.3 Result

In the zero-change control the two separating pattern classes had count zero in all twelve cells and the two separating splits tied exactly in all twelve, and none of the twelve real decision cells failed to compute (Table S5). The real data then miss the bar in both prespecified ways. The statistic is $S = 0.899655$ against the threshold of 3.0, and its two quartet partial sums disagree in sign, -3.612751 and $+4.512406$: one quartet leans toward the published split and the other against it. That the two quartet sums need not agree is the four-taxon counterpart of

Quantity	Value	Requirement	Outcome
Zero-change control: counts of $xyxy$, $xyyx$	0, 0 in 12/12 cells	zero in all cells	pass
Zero-change control: $Z(13\|24) = Z(14\|23)$	identical fractions, 12/12 cells	forced by symmetry	holds
Zero-change control: S	-54.300272	none (reference value)	
Decision cells not computable	0 of 12	≤ 2	not triggered
Real statistic S	0.899655	≥ 3.0	fail
Quartet partial sums	-3.612751 / $+4.512406$	same sign	fail
Damage rung $k = 1$	0/50; planted -54.30 to -49.17	clean iff 0/50	clean
Damage rung $k = 2$	0/50; planted -54.30 to -47.84	clean iff 0/50	clean
Damage rung $k = 5$	0/50; planted -54.06 to -43.33	clean iff 0/50	clean

Table S5: The tortoise test, every prespecified check printed. The construction control passes, the forced tie holds and no cell was lost, so the prespecified decision rules apply; the real data fail both signal requirements (short of the threshold, and sign-discordant between the two quartets); and all 150 damage-planted replicates, in which the cave-bone row is the modern islander's own sequence with deamination-class changes planted, lie more than 44 log units below the real data with zero exceedances, so the cave-bone sequence is not a damaged copy of the modern sequence (for the behaviour of S under both placements see Table S6). S is a sum of twelve exact log Bayes factors (natural logarithms) used only for calibration against the planted null. Scope: one 708-site mitochondrial locus, 14 informative sites, consensus sequences rather than reads, JC69; at 50 replicates, clean means zero exceedances.

the incompatibility among leave-one-out quartets in main-text Section 5. Of the three outcomes, the data give the third: the locus does not support the published placement at this threshold, and, as shown next, the shortfall is not what a damaged copy of the modern sequence produces.

The second finding is narrower than an exclusion of damage. None of the 150 damage-planted replicates scores as high as the real data (zero of fifty at each rung), and the planted statistics span -54.30, the zero-change value, to -43.33, more than 44 log units below the real value. What this excludes is the hypothesis the replicates encode: a cave-bone sequence that is the modern islander's sequence, over the cave bone's coverage, altered only by deamination-class changes at up to five times the published rate. It does not address a cave lineage that is genuinely sister to the modern islander at the divergence the sequences show (18 differences), with or without damage; that hypothesis is calibrated in Section S7.4. At fifty replicates the resolution is itself stated: clean means zero exceedances in fifty, and no finer rate is resolved.

S7.4 Calibration of the decision rule

The rule $S \geq 3.0$ with concordant quartet sums was fixed in the design and calibrated there only against the planted-damage null. Its behaviour under the two competing placements was determined afterwards by simulation. Branch lengths were fitted by maximum likelihood under JC69 to the 687 columns complete in all five sequences, once on the published tree (cave bone sister to the *C. hoodensis–C. abingdonii* pair) and once on the alternative (cave bone sister to the modern islander). The fitted internal branches are short: 0.0049 and 0.0028 expected substitutions per site on the published tree and 0.0015 and 0.0015 on the alternative, that is one to three expected substitutions over the locus, against pendant branches of 0.005 (cave bone), 0.017 (modern islander), 0.013 and 0.009 (sister species) and 0.20 (outgroup); the two fits differ by 1.03 log-likelihood units. Alignments of 708 columns were simulated under JC69 on each fitted tree; the pattern of missing sites of each real sequence was imposed on the corresponding

simulated row, so that every cell drops exactly the columns the real cell drops; and each replicate was cut into the six tiles, scored on the two decision quartets and passed through the rule exactly as the real data were. In further sets of replicates C→T and G→A changes were planted on the simulated cave-bone row at five times the base rate, the internal branches were tripled, or they were set to zero. To carry the uncertainty of the fitted lengths, a further 300 replicates per tree were simulated with all branch lengths drawn anew for each replicate from their joint posterior given the tree and the 687 columns, under the edge prior of main-text Eq. (5). The posterior means of the internal branches, 0.0066 and 0.0042 on the published tree and 0.0030 and 0.0028 on the alternative, exceed the fitted values because the likelihood of a branch carrying one to three changes is skewed toward longer lengths. The 36 marginal likelihoods of each replicate were evaluated by Gauss–Jacobi quadrature in the edge variables of main-text Eq. (5), a floating-point evaluation that reproduces the exact values of the twelve real cells and the twelve zero-change cells to 1.5×10^{-6} log units.

Table S6 gives the result. Under the published placement the rule is passed in 86% of 300 replicates (95% interval 81–89%), and a statistic at or below the observed 0.90 occurs in 6%; under the alternative the rule is passed in 20% (16–25%), and a statistic at or above 0.90 occurs in 28%; on the star tree the rule is passed in 31% and the median statistic is 0.9. Sign discordance between the two quartet sums, observed in the real data, occurs in 6%, 16% and 34% of replicates under the published tree, the alternative and the star tree. Planted damage lowers the two pass rates by two and three to four percentage points and moves the tail fractions at the observed value by three points or less, and tripling the internal branches turns the rule into a sharp one (99% and 5%). Drawing the branch lengths from their posterior instead of fixing them at the fitted values leaves the pass rate under the published placement at 85% (81–89%) and on the star tree at 33%, and lowers it under the alternative to 12% (9–16%), with a statistic at or above 0.90 in 20%. The fitted-length rows are thus the less favorable statement for the rule. The threshold of 3.0, decisive for a single log Bayes factor, is not stringent for a sum of twelve tile terms, whose standard deviation is 4 to 8 log units at these branch lengths; a pass would have been weak confirmation, since the alternative tree passes once in five to once in eight and a star tree once in three. The observed shortfall with discordant signs is mildly atypical of the published placement, typical of a tree whose internal branches are near zero, and compatible with both placements. Deamination-class damage at up to five times the published rate neither produces a spurious pass nor accounts for a shortfall once the cave lineage has a branch of its own; what the planted null of Table S5 excludes is a cave-bone sequence that merely copies the modern islander's.

S7.5 Qualifications

Recoded to purines and pyrimidines, so that only transversions remain visible, the alignment retains no informative site for any ingroup grouping: the entire placement signal is carried by transitions, the substitution class that deamination produces. This was decided by counting before any evidence computation and qualifies every number above. Robustness variants were run under a prespecified rule that they may qualify the outcome but not reverse it. Replacing the holotype fragment by the prespecified alternate cave-bone fragment drives the statistic strongly negative, and replacing the modern islander by an alternative modern record reproduces the primary sign discordance. A variant built on a shorter published control-region record covers only part of the twelve cells and is not commensurable with the others. Every decision cell was computed once; one, computed twice as a check, gave identical values.

True tree	internal branches	damage	n	P(pass)	95% interval	$P(S \geq 0.90)$	P(discordant)	median S (5%, 95%)
published	ML: 0.0049, 0.0028	none	300	0.86	0.81–0.89	0.94	0.06	11.3 (0.0, 25.2)
published	ML	5×	300	0.83	0.79–0.87	0.91	0.07	10.6 (−1.3, 24.9)
published	ML ×3	none	120	0.99	0.95–1.00	0.99	0.01	29.5 (11.5, 61.6)
published	posterior: 0.0066, 0.0042	none	300	0.85	0.81–0.89	0.93	0.07	12.7 (0.4, 36.9)
cave bone with modern	ML: 0.0015, 0.0015	none	300	0.20	0.16–0.25	0.28	0.16	−2.2 (−11.5, 7.0)
cave bone with modern	ML	5×	300	0.16	0.13–0.21	0.29	0.22	−1.6 (−11.3, 6.5)
cave bone with modern	ML	5×, matched count	120	0.16	0.10–0.23	0.34	0.21	−1.5 (−11.5, 6.8)
cave bone with modern	ML ×3	none	120	0.05	0.02–0.10	0.07	0.07	−7.7 (−26.5, 1.7)
cave bone with modern	posterior: 0.0030, 0.0028	none	300	0.12	0.09–0.16	0.20	0.15	−4.7 (−21.2, 6.3)
star	0, 0	none	120	0.31	0.23–0.40	0.49	0.34	0.9 (−4.7, 8.7)
star	0, 0; pendants posterior	none	300	0.33	0.28–0.38	0.53	0.34	1.2 (−5.0, 8.8)
identical copy (design null)	–	none; 1×; 2×; 5×	1; 50; 50; 50	0	–	0	–	−54.3; max −49.2; −47.8; −43.3

Table S6: Calibration of the tortoise decision rule (pass: $S \geq 3.0$ and both quartet sums positive) on alignments simulated under JC69. True tree: the published placement (cave bone sister to the *C. hoodensis*–*C. abingdonii* pair) or the alternative (cave bone sister to the modern islander); branch lengths fitted by maximum likelihood to the 687 complete columns of the five real sequences under each tree (internal branches in expected substitutions per site; pendant branches 0.005 cave bone, 0.017 modern islander, 0.013 and 0.009 sister species, 0.20 outgroup under the published tree, similar under the alternative), or the internal branches tripled, or set to zero; in the rows marked 'posterior' every replicate is simulated at its own draw of all branch lengths from their posterior given the tree (prior of main-text Eq. (5); column two then gives the posterior means of the internal branches). Damage: C→T and G→A changes planted on the simulated cave-bone row at five times the base rate of 0.01 per C or G site (about 17 changes; 'matched count' rescales the rate to the design's 11.75 expected changes). Every replicate has 708 columns, carries the real sequences' pattern of missing sites, and is tiled, scored and judged exactly as the real data; $P(S \geq 0.90)$ is the fraction at or above the observed statistic, P(discordant) the fraction whose two quartet sums differ in sign. Intervals are Wilson 95% binomial intervals. The last row restates the planted-damage null of Table S5, in which the cave-bone row is the modern islander's own sequence.

Neither finding says the published placement is wrong. Nothing here re-places the extinct lineage, and a placement that its locus cannot confirm may still be true. The exact values give a definite two-part statement on the only placement evidence that exists for these bones. There is no support for the placement at the prespecified threshold, and the shortfall is not the signature of a deaminated copy of the modern sequence. The calibration of Section S7.4 adds that one locus of this length, at the branch lengths these sequences support, separates the two placements only weakly, so that the shortfall constrains the placement little in either direction. Where a sampling-based analysis falling short would leave open whether the data or the sampler failed, here the shortfall is a property of the data.

S8 *Anopheles gambiae* complex window scan

Section 6 of the main text uses a rebuilt sliding-window scan of the *Anopheles gambiae* complex. It compares floating-point topology scores with exact values on every window where an exact value could be computed. This section describes the rebuild and its fidelity, the scan, the exactly evaluable subset and the comparison, the provenance of the published data, and the limitations. The scan takes no position on gene flow in the complex; a per-window preferred topology is a support quantity under one substitution model and is never read here as a local species tree.

S8.1 Data and rebuild

The *An. gambiae* species complex comprises the principal African malaria vectors and is among the most closely studied cases of interspecific gene flow in any animal genome (Fontaine et al., 2015; Wen et al., 2016; Thawornwattana et al., 2018). The published analyses turn on a chromosome-level discordance: autosomal windows group *An. arabiensis* with the *An. gambiae–An. coluzzii* pair, whereas X-chromosome windows group it with *An. quadriannulatus*. Fontaine et al. (2015) read the autosomal signal as extensive introgression, and the coalescent reanalyses contest parts of the rooting and interpretation (Wen et al., 2016; Thawornwattana et al., 2018).

The source data are the 2015 release's archived package (Dryad doi:10.5061/dryad.f4114, four files) and the autosomal windows of the 2020 successor archive as published. The window sets of the 2018 coalescent reanalysis could not be obtained as published (Section S8.4). The alignment was therefore reconstructed from the 2015 raw whole-genome alignments following the published methods rule by rule, with thirteen ambiguities in those rules recorded together with the resolution adopted for each. Each window carries twelve ingroup rows and, where the archive provides one, an outgroup row. The rebuild was verified against the surviving successor archive of the same loci. Of that archive's autosomal windows, 98.4% of the noncoding and 98.9% of the coding ones map into the reconstructed alignment with identical ungapped content on the two *An. gambiae* rows. On a 400-locus sample all twelve ingroup rows matched on 399. The published summary statistics reproduce on the rebuild: the relative-rate table to $|\Delta| \leq 0.036$, and the pattern of maximum-likelihood block-tree proportions in both data classes, including the X-versus-autosome switch. The one discrepancy is on the noncoding X, where the rebuilt blocks place the outgroup root on a different edge. The locus counts do not reproduce: the rule-based cutter yields 23% more noncoding and 57% more coding loci than the reanalysis published (57,592 noncoding, 23,505 coding). One selection step of the original extraction scripts, lost with their archive, is not recoverable from the published rules. The reconstruction is therefore a faithful superset rather than the published set, and every X-chromosome window below carries that qualification.

S8.2 The scan

The scan scores 81,113 windows: 75,883 autosomal windows of the successor archive as published (those of its 77,004 with an attachable outgroup row; the 1,121 without one are listed and set aside) and 5,230 X windows from the reconstruction. Each window is scored on three four-species quartets drawn from six ingroup species (*An. gambiae*, *An. coluzzii*, *An. arabiensis*, *An. quadriannulatus*, *An. merus*, *An. melas*), with *An. christyi* as outgroup throughout. The three quartets isolate the three axes of the published debate: the *arabiensis* attachment, the root placement, and the *arabiensis*–(*gambiae*, *coluzzii*) introgression axis. That is 243,339 window–quartet rows, every one scored. Each row carries the three posterior topology probabilities under JC69, with the exponential branch-length prior and uniform topology prior of main-text Eq. (2), computed in floating-point arithmetic. It also carries the alignment's distance below the saturated multinomial model (Goldman, 1993). On 25,362, 15,586 and 33,535 windows of the three quartets a relabeling of the taxa maps the window's pattern counts to themselves. By the argument of main-text Eq. (17) some pair of topologies then has exactly equal evidence under the Jukes–Cantor model with exchangeable branch priors, whatever the method. These windows are flagged before any number is computed.

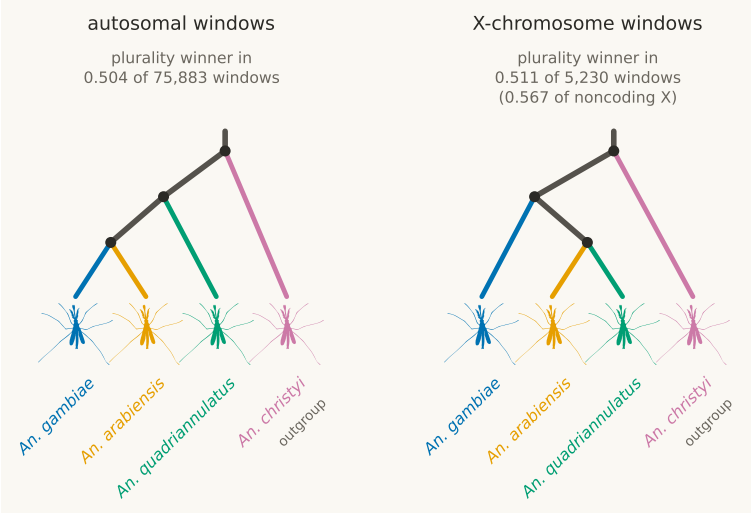

Figure S3: The per-window plurality topology flips between the autosomes and the X chromosome on the *arabiensis*-attachment quartet, drawn rooted on the outgroup *An. christyi*: autosomal windows pair *An. arabiensis* with *An. gambiae* (0.504 of 75,883 windows) and X windows pair it with *An. quadriannulatus* (0.511 of 5,230 windows, 0.567 on the noncoding X), the direction both published interpretations predict. These are window pluralities under JC69, counted with no margin over denominators that include the quartet's 25,362 exactly tied windows; they are not local species trees, no gene-flow inference is drawn, and the X windows carry the reconstruction's superset qualification. The figure is a check on the fidelity of the rebuild.

The known per-window pattern reproduces descriptively on the scored data (Fig. S3). The shares below are pluralities of the preferred topology counted with no margin over all windows; an exactly tied window is assigned there to one of its tied topologies by storage order, which carries no information. Table S7 therefore recomputes the shares by chromosome arm under three rules that involve no tie-breaking: tied windows set aside, tied windows split fractionally among the tied topologies, and the mean over windows of the three posterior topology probabilities. The reversal holds under all three on every autosomal arm, on both autosomal inversion regions and on the X; only the pericentromeric X2 region (1,425 windows), which is close to even, changes its leading topology with the rule. On the *arabiensis*-attachment quartet the plurality flips between autosomes and X in the direction that both published interpretations predict. Of autosomal windows 0.504 pair *arabiensis* with *gambiae* (against 0.212), and of X windows 0.511 (0.567 noncoding) pair it with *quadriannulatus* (against 0.256). On the introgression axis the X is nearly uniform in preference (0.799 for the species grouping, 0.870 noncoding; 0.885 and 0.907 with tied windows set aside) while the autosomes split three ways (0.399, or 0.452 of untied windows, for the species grouping); on the root axis single windows are nearly uninformative against the distant outgroup. We report the flip only as confirmation that the rebuild behaves as the original did.

S8.3 Exact values on the evaluable subset

Two preset limits define the evaluable subset. The first, at most seven mixed sites, the count that governs the cost of exact evaluation (main-text Eq. (14)), selects 7,054 of the 243,339 rows.

Table S7: The autosome-versus-X reversal on the *arabiensis*-attachment quartet does not depend on how exactly tied windows are counted. Each cell gives the share of windows preferring AG|QO (*An. arabiensis* with *An. gambiae*, the autosomal grouping) and then AQ|GO (with *An. quadriannulatus*, the X grouping); the third topology takes the remainder. "Printed": plurality over all windows with ties assigned by storage order, as in Fig. S3. "Set aside": windows whose largest evidence is exactly tied (column 3, certified before any number is computed by the relabelling argument of main-text Eq. (17)) cast no vote; shares are of the remaining windows. "Split": a window tied between k topologies at its maximum gives $1/k$ to each. "Mean posterior": the mean over windows of the posterior topology probabilities under the uniform topology prior. Regions follow Thawornwattana et al. (2018): 2La and 3La are the inversion regions of 2L and 3L, Xag the distal inversion region of the X and X2 its pericentromeric remainder. JC69, floating-point scores; no window has two untied log evidences within 10^{-9} of each other at its maximum. Setting aside in addition every window whose two largest log evidences differ by less than 0.016 (the largest exact separation at which a floating-point preference disagreed with an exact one on the evaluable subset below) or by less than 1 changes no leader outside X2 (autosomes 0.509/0.228 and 0.559/0.202; X 0.218/0.572 and 0.184/0.647).

region	windows	tied at max	printed	set aside	split	mean posterior
2L	19,234	1,878	0.530 / 0.198	0.533 / 0.212	0.515 / 0.223	0.471 / 0.254
2R	23,474	2,522	0.496 / 0.215	0.499 / 0.232	0.483 / 0.242	0.443 / 0.270
3L	14,163	1,402	0.543 / 0.197	0.546 / 0.212	0.527 / 0.223	0.477 / 0.254
3R	19,012	1,924	0.458 / 0.235	0.451 / 0.253	0.440 / 0.260	0.413 / 0.282
autosomes, all	75,883	7,726	0.504 / 0.212	0.505 / 0.229	0.488 / 0.238	0.449 / 0.266
2La	9,446	817	0.579 / 0.159	0.585 / 0.168	0.565 / 0.182	0.507 / 0.225
3La	8,172	776	0.547 / 0.193	0.549 / 0.207	0.531 / 0.218	0.480 / 0.252
X, all	5,230	584	0.256 / 0.511	0.221 / 0.565	0.232 / 0.541	0.259 / 0.491
Xag	3,805	423	0.208 / 0.572	0.172 / 0.634	0.187 / 0.603	0.224 / 0.540
X2	1,425	161	0.384 / 0.348	0.353 / 0.381	0.351 / 0.376	0.351 / 0.358
X noncoding	2,733	152	0.225 / 0.567	0.206 / 0.594	0.212 / 0.581	0.241 / 0.527

The second, at most 150 sites of length, decides which of those were computed. Of the 7,054, 5,375 have at most 150 sites and were computed exactly; the other 1,679 exceed that length. On 2,997 of the 5,375 windows the exact posterior is tied at its maximum, so the exact values define no preferred topology there and a floating-point score can pick one only by rounding error. On the 2,378 windows with a unique exact maximum the floating-point scan selects the same topology on 2,374 and a different one on 4. The largest exact log-posterior separation at any of the four is 0.016 log units (the other three are 5.2×10^{-3}, 1.4×10^{-4} and 9.7×10^{-5}), and across all 5,375 windows the scan's absolute deviation in posterior probability has mean 1.7×10^{-4} (95th percentile 4.8×10^{-4}). Two of the four disagreements are windows on which the scan ranked an exactly tied pair above the exact maximum by less than its own numerical error. It is a measurement on the evaluable subset only, the short, low-complexity corner that holds 2.2% of the rows, and says nothing about the scan's error elsewhere. We know of no published genome scan that reports exact per-window marginal likelihoods, a statement we make only to the extent of our search.

S8.4 Provenance and limitations

The locus set of the 2018 reanalysis (Thawornwattana et al., 2018) was not deposited with that paper, and the server it referenced is no longer online. We were unable to obtain it and therefore reconstructed the loci from the 2015 alignments (Section S8.1). The 2015 release (Dryad doi:10.5061/dryad.f4114) and the 2020 successor archive (an Internet Archive capture; autosomes only, outgroup removed) were used as published.

The limitations are these. No simulation with planted signal was run for this scan, so a failure mode of the floating-point scores confined to long, high-complexity windows would not have been detected. The exact values act as truth only on the evaluable subset. The analysis followed a written plan fixed before any evidence value was computed, but it was not preregistered and no decision rule was fixed in advance in the manner of Section S7. One selection step of the 2018 extraction is unrecovered, the locus counts do not reproduce, and every X window inherits the superset qualification; the reconstruction, though verified, is not the published dataset. About 2.4% of mapped archive loci carry one- or two-base shifts of the window start on non-*gambiae* rows, a convention difference that changed no scored value. The reconstruction and the scored windows have not yet been deposited publicly.

S9 Model choice on short loci

Section 6 of the main text states that a Jukes–Cantor-versus-Kimura log Bayes factor on one alignment carries no estimator error when both marginal likelihoods are evaluated exactly. It quotes a mean preference of 7.99 log units for JC69 over eight draws from the short windows of the mosquito scan. This section gives that comparison in full. It then reports a larger, prespecified test at tRNA length on two quartets, in which the Kimura and general Markov marginal likelihoods had to be estimated by sampling with the software of Schwartz (2026). Throughout, the model evidence of a locus is the evidence marginalized over the three quartet topologies with equal weights, so no topology is chosen; JC69 values are exact, and every comparison that involves a sampled value is an estimate whose uncertainty is the spread of four independent runs.

S9.1 Short windows on which both marginal likelihoods are exact

From the scan of Section S8, fifteen window–quartet pairs, on seven distinct windows, have at most six distinct site patterns and at most 30 usable (gap- and ambiguity-free) sites. Eight of the fifteen were drawn without replacement with a recorded random seed. Two windows were drawn under two quartets each, with identical site-pattern counts under both, so the eight draws cover six distinct windows, of 8 to 29 sites (Table S8). On windows this small the Kimura two-parameter evidence is itself an exact fraction under the branch-wise prior of Section 3.1 of the main text. Under that prior each branch has its own transition and transversion rates, and its two edge variables are uniform on the admissible region of main-text Eq. (9), independently on the five branches. Both marginal likelihoods were stored as integer ratios, and the differences quoted are floating-point views of exact ratios. The stratum is small and deliberately biased toward tiny, conserved windows, and its results describe that stratum only.

Table S8 lists the six windows with both exact values. The mean of $\log \bar{Z}(\mathrm{JC69}) - \log \bar{Z}(\mathrm{K2P})$ over the eight draws is +7.99 log units (standard deviation 1.79, standard error 0.63), with the Kimura prior properly normalized on its admissible region (area 2/3 per branch), as in

Table S8: The six *Anopheles* windows of the Jukes–Cantor versus Kimura comparison (main-text Section 6). Each row is one distinct window of the scan of Section S8; the first column gives its position among the eight seeded draws, and the two windows marked ‡ were drawn twice, once under each of two quartets whose four rows have identical site-pattern counts in that window, so that the two draws have identical values. N is the number of gap- and ambiguity-free sites, K the number of distinct alignment columns (site patterns before reduction to pattern classes) and m the number of mixed sites; $\log \bar{Z}$ is the natural logarithm of the evidence averaged over the three quartet topologies with equal weights, an exact fraction under both models, printed to ten significant digits; $\Delta = \log \bar{Z}(\mathrm{JC69}) - \log \bar{Z}(\mathrm{K2P})$.

draw	region	class	index	quartet	N	K	m	$\log \bar{Z}$(JC69)	$\log \bar{Z}$(K2P)	Δ
1	3R	coding	02303	r	8	4	0	−20.63450347	−24.86642541	4.2319
2, 4‡	2R	noncoding	01946	r, a	28	4	0	−54.21402555	−63.78206024	9.5680
3	3L1	coding	00517	r	23	4	0	−46.33620549	−55.00874343	8.6725
5	2R	noncoding	10626	r	25	6	2	−57.39570884	−65.57268554	8.1770
6, 7‡	3La	coding	01172	i, a	29	5	1	−60.22378979	−68.81352585	8.5897
8	2R	noncoding	01382	r	20	6	2	−49.64962136	−56.17840514	6.5288
mean over the eight draws (standard error)										7.99 (0.63)

All six are autosomal windows of the 2020 archive as published, with the *An. christyi* row taken from the rebuilt alignment of Section S8.1; region, class (coding or noncoding) and index are the fields of the scan's window identifier, the index being the position of the locus in the archive block for that region and class; 3L1 and 3La are the archive's segments of arm 3L. Quartets, each with the outgroup *An. christyi*: a, the *arabiensis*-attachment quartet (*An. arabiensis*, *An. quadriannulatus*, *An. gambiae*); r, the root quartet (*An. merus*, *An. melas*, *An. gambiae*); i, the introgression quartet (*An. gambiae*, *An. coluzzii*, *An. arabiensis*). ‡ Drawn twice: draws 2 and 4 (quartets r and a) and draws 6 and 7 (quartets i and a). The three windows with $m = 0$ consist of constant sites only; in the other three every mixed site is a singleton. On the first five rows the three topologies have identical evidence under both models; on the last, two topologies tie and the third leads them by 0.04 log units under JC69 and 0.01 under K2P. The standard deviation of Δ over the eight draws is 1.79; the exact fractions are records of Supplementary Data S10.

Section 3.1 of the main text. Every draw favors JC69, by 4.23 to 9.57 log units. The standard error is computed over the eight draws as drawn; read against six distinct windows the separation is still many standard errors. One caveat applies to any reuse of these windows for topology. Under both models seven of the eight draws (five of the six distinct windows) are exact three-way topology ties, and on the eighth one topology leads by only 0.04 log units under JC69 and 0.01 under K2P. The stratum ranks models, not topologies, and the trivial agreement of the two models on the preferred topology there says nothing about tree inference.

S9.2 A two-quartet test at tRNA length

Those short windows say which model such tiny loci prefer; the question here is when a richer model earns its additional parameters. Eighteen mitochondrial tRNA genes were scored: Ala, Arg, Asn, Asp, Cys, Gln, Glu, Gly, His, Ile, Leu-CUN, Leu-UUR, Met, Pro, Ser-UCN, Trp, Tyr and Val. Each has roughly 55 to 75 usable sites once the quartet alignments are folded, and each was scored at two divergence depths. The shallow depth is the hominid quartet (human, chimpanzee, gorilla, orangutan); the deep depth is the vertebrate quartet of the tRNA-Gln row of main-text Table 2 (mouse, rat, chicken, *Xenopus*), on this experiment's own alignments. The models compared are JC69, K2P (which adds the transition/transversion distinction and, under the branch-wise prior, five prior dimensions) and the general Markov model (GM), in which every

branch carries its own unconstrained substitution matrix. JC69 values are exact. K2P and GM values at these lengths were estimated with the sampler of Schwartz (2026), four independent runs per value. That sampler had been checked against exact values only on alignments of at most 34 sites, shorter than these loci.

Data and design. The thirty-six locus–quartet alignments come from public mitochondrial genomes. The shallow set is the eighteen hominid quartet alignments of the demonstration in Schwartz (2026), reused unchanged. The deep set was extracted from four GenBank genomes (V00711 mouse, X14848 rat, X52392 chicken, M10217 frog) under a fixed protocol, eighteen of eighteen loci, with no exclusions. Each of the 36 pairs was scored under three models, three topologies and three data variants, the full alignment and the two halves of a random site split fixed in advance. That is 972 evidence cells in all, none of which failed. The hypothesis, the split rule, the selection rule and the scoring rule were fixed in a written preregistration before any split was drawn or any held-out cell scored. The hypothesis was that a per-locus composite, the model selected for each locus on split A, beats every single global model on the held-out split B if and only if the corpus mixes divergence regimes. Before the full computation a positive control passed: the tRNA-Gln pattern counts of Table 2 run through the production evaluator reproduce the three exact values of main-text Table 3 to within 5.0×10^{-8} log units.

The regime mixture. On the shallow quartet JC69 is preferred at sixteen of the eighteen loci on full data, K2P at two and GM at none. On the deep quartet the transition/transversion signal emerges and K2P is preferred at nine, JC69 at nine, GM again at none (Table S9). The full-data difference $\log Z(\mathrm{JC69}) - \log Z(\mathrm{K2P})$ moves from a median of $+6.03$ log units on the shallow quartet to -0.12 on the deep one, and depth moves seventeen of the eighteen loci toward K2P. The exception, Asn, moves the other way by 4.2 log units. Two caveats attach to the deep full-data numbers. The deep K2P cells have 55 to 74 usable sites, beyond the lengths at which the K2P sampler was checked against exact values, so their run-to-run spreads do not include any systematic error of the sampler. And the corpus-level deep advantage is thin: deep K2P at every locus beats deep JC69 at every locus by 0.8 log units in total. The K2P preference of 3.8 log units reported for the tRNA-Gln alignment in Schwartz (2026) is therefore locus-specific; on this experiment's own Gln alignment the preference is 5.22 log units.

The held-out test. Selection on split A picked K2P at three of the thirty-six pairs (Tyr deep by $+8.45$ log units, Ser-UCN deep by $+0.95$, Glu shallow by $+0.21$) and JC69 at the other thirty-three, with no ties; GM was last everywhere. Of the nine deep loci that prefer K2P on full data only two kept that preference on the half-length selection split. Split B was then scored in two ways, which answer different questions (Table S10).

The score fixed in the preregistration is the evidence of split B on its own, $\sum \log \bar{Z}_M(B)$ with M the model selected on A, each $\bar{Z}$ averaged over the three topologies with equal weights as elsewhere in this section. It asks whether a model chosen by its evidence on one half is the model the evidence prefers on an independent half of the same length. It was not: all three K2P picks reversed on B (-3.03, -1.83 and -5.43 log units against JC69, the $+8.45$ selection included), and these reversals are the whole of the margin. Under this score the composite loses to JC69-at-every-locus by 10.295 ± 0.0008 log units, while beating K2P-at-every-locus by 113.189 ± 0.004 and GM-at-every-locus by 2055.6 ± 0.4. The JC69 leg is exact, so the first $\pm$ is the sampling spread of three K2P cells of 31, 34 and 34 sites, at the lengths where the sampler

Table S9: Divergence depth moves seventeen of the eighteen loci toward K2P; **Asn** (bold) is the one exception. Each row gives $\Delta = \log Z(\mathrm{JC}) - \log Z(\mathrm{K2P})$ for one locus on full data, marginalized over the three topologies with equal weights, under the branch-length prior of main-text Eq. (6) and the branch-wise Kimura prior fixed in the preregistration; positive favors JC69. The depth shift is shallow minus deep, so positive means depth moves the locus toward K2P. The shallow median is +6.03 log units and the deep median −0.12; on deep data K2P is preferred at nine loci and JC69 at nine. JC69 values are exact and K2P values are sampling estimates whose run-to-run spread is at most 0.003 log units, below the last displayed digit; the K2P cells have 55 to 75 usable sites, beyond the lengths at which the sampler was checked against exact values, so that spread does not include any systematic error.

locus	shallow Δ (log units)	deep Δ (log units)	depth shift (log units)
Ala	+9.38	+0.72	+8.66
Arg	+8.67	+3.31	+5.36
Asn	**+1.50**	**+5.70**	**−4.20**
Asp	−0.54	−3.53	+3.00
Cys	+7.81	+0.65	+7.17
Gln	+5.96	−5.22	+11.18
Glu	−0.77	−0.90	+0.13
Gly	+4.77	+1.60	+3.17
His	+6.23	−3.29	+9.51
Ile	+5.90	−3.47	+9.36
Leu-CUN	+12.06	+7.89	+4.17
Leu-UUR	+6.11	+3.68	+2.43
Met	+10.42	+7.54	+2.87
Pro	+6.81	+1.51	+5.30
Ser-UCN	+6.92	−4.23	+11.15
Trp	+2.10	−5.83	+7.93
Tyr	+1.32	−4.49	+5.81
Val	+5.46	−2.46	+7.93

was checked.

The second score is predictive: $\log p_M(B \mid A) = \log \bar{Z}_M(A \cup B) - \log \bar{Z}_M(A)$, the probability of the held-out half given the selection half, in which the posterior from A over branch parameters and topologies is the prior under which B is predicted ($\bar{Z}_M(A \cup B)/\bar{Z}_M(A) = \sum_\tau p_M(\tau \mid A)\, p_M(B \mid A, \tau)$, so the equal topology weights enter only through A). It asks which model, having seen A, predicts B better, and it penalizes a model for its prior volume once rather than twice. For each pair and model the two scores differ by $\log \bar{Z}_M(A \cup B) - \log \bar{Z}_M(A) - \log \bar{Z}_M(B)$, the amount by which A sharpens the prediction of B; here that amount is 3.3 to 11.3 log units under JC69 and 3.5 to 18.5 under K2P, larger under K2P at 35 of the 36 pairs (median difference 5.7). Under the predictive score the composite and JC69-at-every-locus are indistinguishable, at -0.117 ± 0.0014 log units (Glu shallow +0.56, Ser-UCN deep +3.28, Tyr deep −3.96), but the ranking of the single-model baselines reverses: K2P-at-every-locus predicts the held-out halves better than JC69-at-every-locus by 56.08 ± 0.015 log units in total, 8.07 on the shallow quartet and 48.01 on the deep one, and better at 25 of the 36 pairs (12 shallow, 13 deep). The composite therefore loses to K2P-at-every-locus by 56.20 ± 0.015 and beats GM-at-every-locus by

Table S10: The held-out half scored two ways. Totals over the 36 locus–quartet pairs of the topology-averaged log evidence of split B on its own (the preregistered score) and of the predictive score $\log p_M(B \mid A) = \log \bar{Z}_M(A \cup B) - \log \bar{Z}_M(A)$; the composite uses at each pair the model selected on split A (K2P at three pairs, JC69 at thirty-three). The lower block gives composite minus each single-model policy; only pairs at which the two policies use different models contribute. JC69 values are exact; ± is the run-to-run sampling spread of the K2P and GM cells (four runs per cell, propagated in quadrature), and for the predictive column it includes the full-data K2P and GM cells of 55 to 75 sites, beyond the lengths at which those samplers were checked against exact values. Under the first score the best single policy is JC69 everywhere; under the second it is K2P everywhere; the composite beats neither.

	$\sum \log \bar{Z}_M(B)$	$\sum \log p_M(B \mid A)$
Composite (selected on A)	-3529.250 ± 0.001	-3240.346 ± 0.001
JC69 at every locus	-3518.955 (exact)	-3240.229 (exact)
K2P at every locus	-3642.440 ± 0.004	-3184.150 ± 0.015
GM at every locus	-5584.87 ± 0.40	-4001.89 ± 0.90
Composite − JC69	-10.295 ± 0.001	-0.117 ± 0.001
Composite − K2P	$+113.189 \pm 0.004$	-56.196 ± 0.015
Composite − GM	$+2055.6 \pm 0.4$	$+761.5 \pm 0.9$
K2P − JC69	-123.484 ± 0.004	$+56.078 \pm 0.015$

761.5 ± 0.9. This comparison uses the full-data K2P cells of 55 to 75 sites, beyond the lengths at which the K2P sampler was checked against exact values, and its ± is run-to-run spread only. Two facts bound the error not covered by that spread: the estimator is an importance-sampling average, whose logarithm errs downward in expectation, and the same estimator under JC69 reproduces exact values at 67 to 69 sites to a mean absolute deviation of 5×10^{-4} log units; erasing the K2P advantage would take a same-signed error of 1.5 log units at every locus.

Neither score is the tautological one. Scored without the split, the same composite shows an advantage of +34.7 log units over JC69 at every locus; the same data select and score, and the number carries no evidential weight.

The preregistered split is one draw. To separate split-to-split variability from sampling error, eight further random site splits were drawn with a recorded seed under the same rule and scored with the same programs and settings at all thirty-six pairs under all three models (Table S11). Selection on the new halves picked K2P at three to eight pairs per split, forty-six picks in all, forty-five of them on the deep quartet, spread over thirteen pairs with no pair chosen on more than six of the eight splits (Trp deep on six; Asp, Cys, Ser-UCN, Tyr and Val deep on five). Twenty-three pairs were never chosen on any split, and GM was never selected and never led JC69 on a selection half. The evidence preference between JC69 and K2P on one half agreed in sign with the other half's at 248 of the 324 pair–splits, which is the number expected (247.8) if the two halves' preferences were independent at the observed rates. Under the evidence score forty of the forty-six K2P picks reversed on the scoring half, and the composite lost to JC69-at-every-locus on all eight new splits, by 4.42 to 19.31 log units (mean 11.88, standard deviation across splits 5.36); the preregistered split's 10.29 is the median of the nine. Under the predictive score the composite led JC69-at-every-locus on six of the eight (by −3.70 to +13.65, mean +6.96, standard deviation 6.71) and lost to K2P-at-every-locus on all eight (by 25.0 to 46.9). On no split, under either score, did the composite beat the better of the two single-model policies. The reversal between the scores is itself stable: K2P-at-every-locus trails JC69-at-every-locus by 133.5 to 158.0 log units under the evidence score and leads it by 29.8 to 52.1 under the

Table S11: Split-to-split variability of the held-out comparison. The preregistered split (row 0) and eight further random site splits of the same 36 locus–quartet pairs, each scored both ways as in Table S10: totals over the 36 pairs of composite minus JC69-at-every-locus, composite minus K2P-at-every-locus, and K2P minus JC69 at every locus, in log units. "K2P picks" is the number of pairs at which selection on half A chose K2P (JC69 at the rest; GM at none). JC69 values are exact; the run-to-run sampling spread of every entry is below 0.01 log units under the evidence score and below 0.04 under the predictive score. The composite is below the better single-model policy in every row under both scores, and the sign of K2P minus JC69 reverses between the scores in every row.

		$\sum \log \bar{Z}_M(B)$			$\sum \log p_M(B \mid A)$		
Split	K2P picks	Comp.−JC69	Comp.−K2P	K2P−JC69	Comp.−JC69	Comp.−K2P	K2P−JC69
0 (preregistered)	3	−10.29	+113.19	−123.48	−0.12	−56.20	+56.08
1	8	−18.29	+131.41	−149.70	+7.82	−28.62	+36.45
2	3	−14.85	+133.09	−147.94	−3.70	−40.36	+36.65
3	8	−19.31	+133.59	−152.90	−1.99	−31.83	+29.84
4	5	−7.10	+150.92	−158.02	+13.65	−24.97	+38.63
5	6	−9.28	+146.13	−155.41	+12.83	−25.18	+38.00
6	6	−13.01	+127.75	−140.76	+9.15	−40.40	+49.55
7	5	−4.42	+144.02	−148.44	+12.71	−25.37	+38.07
8	5	−8.78	+124.72	−133.50	+5.21	−46.88	+52.08
Mean (1–8)	5.75	−11.88	+136.45	−148.34	+6.96	−32.95	+39.91
SD (1–8)	1.67	5.36	9.40	7.96	6.71	8.50	7.31

predictive score, on every new split, predicting the held-out half better at 20 to 26 of the 36 pairs. That lead is 18.1 to 42.3 log units on the deep quartet on every split and −5.7 to +24.0 on the shallow one, and the preregistered split's 123.5 and 56.1 make it the most favorable of the nine to K2P under both scores. Within a split the sampling spread of any of these differences is at most 0.04 log units, two orders of magnitude below the spread across splits.

The conclusion, with its scope: at loci of about 55 to 75 sites, choosing a substitution model locus by locus from its evidence on half the sites gave no advantage under either score, although the regime mixture it targets was real and measured. Under the evidence score the choice did not replicate on the other half and the composite lost to JC69 everywhere; under the predictive score the composite tied JC69 everywhere and lost to K2P everywhere, the model the selection rule had rejected at thirty-three of thirty-six pairs. The second comparison says something the first cannot: at these lengths the evidence prefers JC69 because the branch-wise Kimura prior of main-text Eq. (9) adds five parameters whose prior volume the information in half a tRNA gene does not outweigh, not because JC69 predicts held-out sites better, which on these data it does not. Which score is the right one depends on the use: the evidence ranks models as prior predictive machines for a whole locus, the predictive score ranks them after part of the locus has been seen, and a reader choosing a model for loci of this size should know that the two disagree here. This is a statement about mitochondrial tRNA genes on two quartets, one model menu and one prior family, not a general claim about model selection, other taxa, longer loci, concatenations, other menus or other priors on the Kimura parameters. The topology preferences in the underlying tables are descriptive and no topology claim is made. The shallow values were known before the design was fixed, and the same authors fixed the design and computed the deep values. The protections are the prespecified rules, the disclosed pilot and the held-out split, which an independent re-derivation of the thirty-six splits reproduced (10.295

log units); the predictive score and the eight further splits are later additions and not part of the preregistered design.

S10 Exact values in machine-readable form

The exact values behind main-text Tables 2, 3, 4 and B1, and those of Table S8, are distributed as Supplementary Data S10. It consists of two files, `S10_exact_values.json` and `S10_exact_values.csv`, with their SHA-256 digests in `S10_exact_values.sha256`, written by the included programs directly from the stored results of the computations. This section says what the files contain and gives, in print, the site-pattern counts of every four-taxon row and the fifteen five-taxon values to ten decimals.

S10.1 Contents of the data file

The file has 84 records: one per alignment, prior and topology for the main-text tables, and one per window and model for Table S8. The three topologies of each of the twelve tRNA quartets of Table 2 give 36 records, and those of the lysozyme, 5S rRNA, twenty-site and two 900-site alignments of Table 3 give 15. The two lines of Table B1 that repeat the tRNA-Phe quartet at prior means 0.1 and 0.01 give six. The fifteen topologies of the hominoid quintet of Table 4 give 15. The twelve records of Table S8 give, for each of the six windows, the Jukes–Cantor and the Kimura evidence averaged over the three topologies with equal weights, as reduced fractions of 21 to 165 digits over 29 to 195 digits with $\log \bar{Z}$ to at least 49 significant digits, together with the unfolded site-pattern counts that are the complete input under Kimura's model and the prior of main-text Eq. (9); these twelve values were computed in rational arithmetic throughout. The 51 records of the four-taxon rows of Table B1 carry the exact value as two decimal integer strings, the numerator and denominator of the reduced fraction $Z = p/q$ of main-text Eq. (6). These run from 48 to 1,320 digits over 87 to 1,682 digits, and each record also gives $\log Z$ as a decimal string of 40 to 50 significant digits. For the twelve tRNA quartets alone the integers run from 131 to 331 digits. The fifteen quintet records carry $\log Z$ to 60 digits and the digit counts of numerator and denominator (295 to 298 over 378 to 381); the quintet integers themselves are omitted. The six records of the two 900-site alignments are marked not exact and carry the floating-point Gauss–Legendre values of Table 3 at every quadrature grid that was run. Each record of the main-text tables also states the model and the prior in both common parameterizations: independent exponential priors on the $2n-3$ branch lengths with rate parameter $4/3$ (mean 0.75 expected substitutions per site), equivalently $\alpha = 1$ with each x_e uniform on $[0, 1]$; the six records at smaller prior means state rate 10 or 100, equivalently $\alpha = 15/2$ or 75 with density $\alpha x_e^{\alpha-1}$. It gives N, the mixed-site count m in the 12|34 frame and for the record's own topology, the taxa in the order that defines the split codes, and the sequence accessions or Rfam seed-row identifiers. It gives the fifteen site-pattern counts in the class order of main-text Table A1, which are the complete input to Algorithm 1. It records how the value was checked: the number of primes in each of the two reconstruction sets and the agreement with floating-point quadrature. And it gives the name and SHA-256 digest of the result file from which each record was taken. The twelve window records and the six records at smaller prior means state their method (rational arithmetic throughout) in place of the prime counts. Two of the four-taxon alignments contain an exact tie, and the file shows it as it is. The 12|34 and 14|23 records of the tRNA-Phe quartet hold identical integer strings, as do the 13|24 and 14|23 records of the twenty-site case, and the six tied pairs of the quintet are marked by their partner. Two further coincidences

among the denominators are genuine and not assembly errors: the plants $Z(13|24)$ and cross-domain $Z(12|34)$ denominators are the same 316-digit integer, and the cross-domain $Z(13|24)$ and $Z(14|23)$ denominators, under distinct numerators, equal exactly twice that integer. The comma-separated file holds the same records with one column per field and per pattern class.

S10.2 Pattern counts and sources

Table S12 gives the site-pattern counts y_k of every alignment of main-text Table B1, in the class order of main-text Table A1; each row is the complete input to Algorithm 1, which returns the corresponding row of Table B1. Rows 16 and 17 are the two 900-site alignments of main-text Table 3, for which only quadrature values exist. Table S13 gives the nine distinct five-taxon values of main-text Table 4 to ten decimals with the digit counts of the reduced fractions.

Table S12: Site-pattern counts y_k for the rows of main-text Table B1 (rows 1–15) and for the two 900-site alignments (rows 16, 17), in the class order of main-text Table A1; $N = \sum_k y_k$.

	N	xxxx	xxyy	xyxy	xyyx	xyyy	xyxx	xxyx	xxxy	xxyz	xyzz	xyxz	xyzx	xyyz	xyzy	xyzw
1	69	59	0	0	0	1	3	1	5	0	0	0	0	0	0	0
2	70	57	2	0	1	1	0	2	6	0	0	0	1	0	0	0
3	75	68	0	1	0	0	2	1	3	0	0	0	0	0	0	0
4	68	35	8	1	0	6	0	5	6	0	4	0	0	0	3	0
5	67	19	0	10	0	7	0	11	0	0	0	1	1	0	18	0
6	66	13	2	1	8	1	7	5	6	3	2	1	5	3	2	7
7	70	20	5	0	4	5	3	7	5	2	4	8	0	6	0	1
8	71	16	2	6	2	7	7	0	5	7	3	5	3	3	4	1
9	63	3	1	5	1	6	5	1	3	2	6	2	11	4	5	8
10	70	20	0	12	1	0	5	2	7	3	2	0	8	3	3	4
11	70	24	2	2	6	2	5	6	6	0	1	4	5	3	2	2
12	72	23	1	8	3	6	5	3	4	4	2	1	4	3	2	3
13	390	344	7	1	0	4	3	14	15	2	0	0	0	0	0	0
14	116	78	0	1	1	2	4	3	24	1	0	1	0	1	0	0
15	20	10	4	2	2	0	0	0	1	0	0	0	0	0	0	1
16[d]	895	677	17	9	13	23	29	33	85	4	0	3	0	2	0	0
17[d]	890	487	46	29	12	52	45	59	69	24	6	18	14	12	13	4

Sources. Rows 1–3: mitochondrial genomes NC_012920 (human), NC_001643 (chimpanzee), NC_001645 (gorilla), NC_001646 (orangutan), the named tRNA gene of each, excised at the coordinates of the annotated tRNA feature; the four sequences were aligned pairwise to a reference among them by the Needleman–Wunsch algorithm (match 1, mismatch -1, gap -2) and the pairwise alignments merged, as in Section 4 of the main text, before columns with a gap were removed. The hominoid quintet of Table S13 adds the tRNA-Phe gene of NC_002082 (gibbon) under the same procedure. Row 4: Rfam RF00005 seed rows V00711.1/3842–3772, X14848.1/3820–3750, X52392.1/5172–5102, M10217.1/5909–5841. Rows 5–12: Rfam RF00005 seed rows, in taxon order, (5) AC004941.2/32735–32806, M22657.1/587–656, M22650.1/459–527, M22655.1/587–657; (6) AF076356.1/60–128, AB049357.1/1–68, AJ428514.1/11535–11602, M10217.1/11492–11561; (7) AF008220.1/5322–5394, DQ023210.1/1–73, X17113.1/351–433, X51824.1/190–272; (8) J01390.1/11761–11831, K00144.1/1–72, AJ223323.1/613–684, K00344.1/1–73; (9) L76657.1/1–67, K03317.1/4–76, AB009835.1/1–71, X13975.1/2575–2637; (10) AB017063.1/58819–58900, M16897.1/160–88, AB027572.1/4261–4342, L02941.1/716–643; (11) DQ023210.1/1–73, AC004941.2/32735–32806, X00083.1/99–171, AJ223323.1/613–684; (12) M87833.1/2781–2865, K02528.1/1–74, X00083.1/99–171, X68198.1/9818–9745. [b] Rows 10 and 11 pair non-orthologous tRNAs (nuclear with organellar genes; Section 4 of the main text); their values describe those alignments, not the species tree. Row 13[c]: the four-species subset (human, gibbon, colobus, marmoset) of the seven-species, 390-bp lysozyme alignment of Messier and Stewart (1997) as distributed with PAML (`lysozymeSmall` data set). Row 14: Rfam RF00001 seed rows X71804.1/109–227 (mouse), X00190.1/1–118 (chicken), J01009.1/607–725 (*Xenopus*), M25016.1/1–119 (*Drosophila*). Row 15: synthetic counts, no sequences. [d] Rows 16 and 17: the human, chimpanzee, gorilla, orangutan subset of the 895-bp alignment of Brown et al. (1982) distributed with PAML, and the *Homo*, *Macaca*, *Saimiri*, *Lemur* subset of the 898-bp alignment of Hayasaka et al. (1988) distributed with MrBayes (8 ambiguous columns removed); no exact value exists for these two rows.

Table S13: The nine distinct exact log marginal likelihoods among the fifteen five-taxon topologies of the hominoid tRNA-Phe quintet ($N = 67$, $m = 13$), $\log Z$ (natural logarithm) rounded to ten decimals, JC69, independent exponential branch-length priors of rate 4/3 (mean 0.75 expected substitutions per site). H human, C chimpanzee, G gorilla, O orangutan, Gi gibbon; $((a, b), e, (c, d))$ has cherries (a, b) and (c, d); the partner is the topology with the identical fraction under the human–gorilla exchange; Δ is the log Bayes factor in favor of row 1; p/q the digit counts of numerator and denominator.

	topology	tied partner	$\log Z$	Δ	digits p/q
1	((H,C),G,(O,Gi))	((C,G),H,(O,Gi))	−189.6451165484	0	296/378
2	((H,G),C,(O,Gi))	–	−190.2076498014	0.5625	298/381
3	((C,O),H,(G,Gi))	((H,Gi),G,(C,O))	−191.3674403814	1.7223	295/378
4	((H,C),O,(G,Gi))	((H,Gi),O,(C,G))	−191.4518420533	1.8067	297/380
5	((H,O),C,(G,Gi))	((H,Gi),C,(G,O))	−191.5212229141	1.8761	297/381
6	((H,C),Gi,(G,O))	((H,O),Gi,(C,G))	−193.6282837068	3.9832	297/381
7	((H,G),Gi,(C,O))	–	−194.0873426113	4.4422	297/381
8	((C,Gi),H,(G,O))	((H,O),G,(C,Gi))	−194.3171363625	4.6720	295/379
9	((H,G),O,(C,Gi))	–	−194.5378507972	4.8927	295/380

S11 Relation to string amplitudes

The Discussion of the main text cites Sturmfels and Telen (2021) on likelihood integrals and scattering amplitudes. The connection is recorded here for the interested reader; nothing else in the paper depends on it. Under $x_e = t_e/(1 + t_e)$ the evidence of main-text Eq. (6) takes the form of the stringy integrals of Arkani-Hamed, He and Lam (2021): powers of the letters t_e, $1+t_e$ and the pattern polynomials, integrated over the positive orthant. It sits in the opposite exponent chamber, however. The data counts enter with the opposite sign: there the polynomials carry negative powers and the integral converges only in a chamber of exponent space, whereas here they carry non-negative powers and the integral converges for every count vector $\mathbf{y} \geq 0$ and every $\alpha > 0$. For integer counts, $\alpha^{-5}Z$ is a rational function of the prior exponents, one exponent α_e per edge if these are allowed to differ, whose only poles lie at $\alpha_e \in \{0, -1, \ldots, -N\}$ (Eq. (8)). As a function of continuous counts at fixed $\alpha > 0$ the evidence is analytic on a neighbourhood of the orthant $\mathbf{y} \geq 0$ and entire in the constant-site count, since $p_{xxxx} \geq 4^{-4}$ on the closed cube. It is not entire in the other counts, because every other pattern polynomial vanishes identically on a face of the cube of codimension two or three (for instance $p_{xxxy} \equiv 0$ on $x_3 = x_4 = 1$, where leaves 3 and 4 cannot differ), so that $\int p_k^s \, \mathrm{d}x$ diverges at a finite negative s. In the shared combinatorics the three quartet topologies are the s-, t- and u-channels of a four-point amplitude, the three ways two incoming and two outgoing particles can pair up. They are equally the three maximal cones of the tropical Grassmannian, which is the space of phylogenetic trees (Speyer and Sturmfels, 2004; Pachter and Sturmfels, 2004). Give the internal edge its own prior exponent α_5. The unnormalized integral $\alpha_5^{-1}Z$ has a simple pole at $\alpha_5 = 0$, from the monomials free of x_5, while the evidence itself, which carries the prior's normalizing factor α_5, remains finite: as $\alpha_5 \to 0$ the prior $\alpha_5 x_5^{\alpha_5 - 1}$ concentrates at $x_5 = 0$, an infinitely long internal branch, and Z tends to the likelihood with that branch saturated, integrated over the four pendant branches. Across a saturated edge the two cherries are independent, so this limit, which is also the residue of $\alpha_5^{-1}Z$, is the product of the evidences of the two two-leaf trees $\{1, 2\}$ and $\{3, 4\}$; we verified the identity in exact arithmetic on a six-site example. This is the analog of

factorization on a massless pole, and here a consequence of Markov independence across a fully randomized edge. The parallel has limits: an amplitude sums its channels, Bayesian selection takes their ratios, and the Bayes factor has no single-amplitude analog. Sturmfels and Telen (2021) observed that open-string amplitudes arise as limits of marginal-likelihood integrals of positive models; the exact values of this paper live in the interior of that interpolation, away from the limiting endpoints. Earlier physics formulations of phylogenetics concern the substitution process itself rather than the parameter integrals (Jarvis et al., 2005).

S12 Controls for the hominoid tRNA-Phe comparisons

Sections 4.2 and 5.2 of the main text report that on the hominid tRNA-Phe quartet the grouping of human with gorilla has the largest evidence by 0.0013 log units although no column supports it, and that the same grouping ranks third once gibbon is added. This section locates the four-taxon margin in the likelihood surfaces and the branch-length prior, and separates the three changes between the four-taxon and the five-taxon comparison.

S12.1 The four-taxon margin: likelihood surfaces and priors

The hominid tRNA-Phe quartet has no parsimony-informative column (59 constant columns and 10 singletons), yet under the model and prior of the main text the topology joining human with gorilla has the largest evidence, by 0.0013 log units. The likelihood was evaluated in floating point by the pruning algorithm; maxima were found by bounded quasi-Newton search from ten starting points, and integrals over branch lengths by Gauss–Jacobi quadrature with 35 nodes per branch, which is exact for these polynomial integrands apart from rounding and reproduces the exact evidences of main-text Table B1 to 10^{-12} log units.

Maximized over all five branch lengths, the three log-likelihoods are equal, -147.3710246 ($N = 69$; -144.2961012 on the quintet's 67 columns): for each topology the maximum lies at internal length zero, where the three trees are the same star tree, with pendant lengths 0.0146 (human), 0.0448 (chimpanzee), 0.0146 (gorilla) and 0.0762 (orangutan), and the derivative of the log-likelihood in the internal length there is -65.5 per substitution per site for all three. Fit therefore does not distinguish the topologies. Table S14 follows the three likelihood functions away from the star tree in two ways. The profile log-likelihood, in which the four pendant lengths are set to their best values at each internal length t_5, is slightly lower for the (human, gorilla) topology than for the other two up to $t_5 \approx 0.59$ (by 0.0018 at $t_5 = 0.1$) and higher beyond, where all three are more than 30 log units below the maximum. The pendant-integrated likelihood $M_T(t_5) = \int L_T \prod_{e \neq 5} \pi(t_e)\, \mathrm{d}t_e$, whose integral against $\pi(t_5)$ is the evidence, is instead higher for the (human, gorilla) topology at every $t_5 > 0$, by 0.0011 log units at $t_5 = 0.02$, 0.028 at 0.1 and 0.44 at 0.5, almost independently of the prior mean on the pendant branches. Near the star tree the difference grows as $c\,t_5^2$ with $c \approx 2.7$.

The evidence margin is the average of this difference over the posterior distribution of t_5. Since the likelihood falls off the star tree as $e^{-\lambda_L t_5}$ with $\lambda_L \approx 65.5$ and the prior is exponential with rate λ, that posterior is close to exponential with rate $\lambda_L + \lambda$, and the margin is approximately $2c/(\lambda_L + \lambda)^2$: 0.0012, 0.0009 and 0.0002 log units at prior means 0.75, 0.1 and 0.01 ($\lambda = 4/3$, 10, 100), against exact values 0.0013, 0.0010 and 0.0002. Half of the margin comes from internal lengths below 0.05 and nine tenths from lengths below 0.09 at prior mean 0.75. The margin thus falls with the prior mean far more slowly than in proportion (by 0.77 rather than 0.13 from mean 0.75 to 0.1), because at those means the posterior of the internal length

is governed by the likelihood; only at mean 0.01 is the prior the narrower of the two. It is an interaction between the likelihood, which differs between topologies only at positive internal lengths and only after the pendant lengths are integrated, and the prior, which sets the weight those lengths receive.

Other prior families confirm this (values on the quartet's 69 columns; Table S15 gives the corresponding values on the quintet's 67 columns). Under gamma priors of shape two on all five branches the margin is 0.0199 log units at mean 0.75 and 0.0151 at mean 0.1 (exact; a gamma prior of integer shape keeps the evidence rational), larger than under the exponential because less prior mass lies near the star tree. Under the mixture prior that the five-taxon model induces on the quartet (Section S12.2) it is 0.0015. Priors that are not exchangeable across branches can reverse it: a gamma branch of shape two on the chimpanzee lineage alone gives −0.0042, on the orangutan lineage −0.0001, whereas on the human, gorilla or internal branch it gives +0.0064, +0.0119 and +0.0041. The exact tie of the other two topologies holds under every exchangeable prior in this list and fails, as it must, only under the two single-lineage priors that the human–gorilla exchange maps onto each other.

Table S14: The hominid tRNA-Phe quartet ($N = 69$, JC69) away from the star tree. Profile log-likelihood (pendant lengths maximized) and pendant-integrated log-likelihood $\log M_T(t_5)$ (pendant lengths integrated against exponential priors of rate 4/3) at fixed internal length t_5, for $T = (\mathrm{H},\mathrm{C}) \mid (\mathrm{G},\mathrm{O})$ and $T = (\mathrm{H},\mathrm{G}) \mid (\mathrm{C},\mathrm{O})$; $(\mathrm{H},\mathrm{O}) \mid (\mathrm{C},\mathrm{G})$ equals the former on both surfaces. Δ, the (H,G) column minus the (H,C) column. At $t_5 = 0$ the three trees coincide and both surfaces take their maxima.

	profile $\log L$			pendant-integrated $\log M$		
t_5	HC|GO	HG|CO	Δ	HC|GO	HG|CO	Δ
0	−147.3710	−147.3710	0	−158.0497	−158.0497	0
0.02	−148.6761	−148.6764	−0.0003	−159.3271	−159.3260	+0.0011
0.04	−149.9713	−149.9720	−0.0006	−160.5973	−160.5927	+0.0046
0.1	−153.7973	−153.7991	−0.0018	−164.3624	−164.3344	+0.0280
0.2	−159.9688	−159.9732	−0.0044	−170.4715	−170.3708	+0.1007
0.3	−165.8739	−165.8819	−0.0080	−176.3518	−176.1503	+0.2015
0.5	−176.8467	−176.8646	−0.0179	−187.3505	−186.9112	+0.4393
0.6	−181.9272	−181.9238	+0.0034	−192.4407	−191.8806	+0.5601
1.0	−199.5269	−199.2078	+0.3192	−209.9016	−208.9357	+0.9658

S12.2 The quartet inside the quintet

The four-taxon and five-taxon tRNA-Phe alignments of the main text are one alignment: the first four rows of the five-taxon star alignment (73 columns before gap removal) are the quartet alignment, and the two differ only in the gap rule, which keeps the 69 columns free of gaps in the four hominids or the 67 free of gaps in all five hominoids. The two columns present at four taxa and absent at five are constant among the hominids (gibbon has a gap in each). Three things therefore change between panels (a) and (b) of main-text Fig. 5: two constant columns are dropped, the prior on the branch lengths of the four-taxon subtree changes, and gibbon's sequence is added. Table S15 separates them.

Deleting gibbon from the five-taxon model changes the prior seen by the quartet. Summing the five-taxon pattern probabilities over gibbon's state removes gibbon's pendant branch, and

the node at which it was attached then joins two branches in series, whose lengths add (the Chapman–Kolmogorov property of the substitution process). The quartet obtained from a five-taxon topology therefore has four exponential branches of rate 4/3 and one branch, the one on which gibbon sat, whose length is the sum of two such exponentials, a gamma variable of shape two (density $-\ln x_e$ in the variable $x_e = e^{-4t_e/3}$). Since gibbon sits on each of the five quartet branches in exactly one of the five topologies containing a given split, the induced prior is the equal-weight mixture of the five assignments of the gamma branch. The moments $\int_0^1 x^j(-\ln x)\,\mathrm{d}x = (j+1)^{-2}$ are rational, so these evidences are again exact rational numbers, computed by the program of Section S3 with the moment map of one branch replaced.

On the quintet's 67 columns the quartet margin for (human, gorilla) under the exponential prior is 0.0014 log units (0.0013 on 69 columns), and under the induced prior 0.0016; in both the two other topologies are tied as identical fractions. The five components of the mixture are informative on their own: placing the gamma branch on the human or the gorilla lineage raises the margin to 0.0067 and 0.0123, on the internal branch to 0.0044, whereas placing it on the orangutan lineage removes it (−0.00001) and on the chimpanzee lineage reverses it (−0.0043). With gibbon's sequence included and the five-taxon evidence summed over gibbon's five placements, the margin becomes −0.3518 log units. The whole of the reversal is thus attributable to gibbon's sequence, which is $e^{0.353} = 1.42$ times more probable under the five extensions of (H, C) | (G, O) than under those of (H, G) | (C, O); the columns and the prior together move the margin by 0.0003 log units. The main text's 0.5625 compares instead the best single topology containing each cherry, the three resolutions of the African-ape trichotomy beside the (orangutan, gibbon) pair.

References

Allman, E. S. and Rhodes, J. A. (2006). The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *J. Comput. Biol.*, 13:1101–1113.

Anderson, S. et al. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, 290:457–465.

Arkani-Hamed, N., He, S., and Lam, T. (2021). Stringy canonical forms. *JHEP*, 02:069. arXiv:1912.08707.

Borinsky, M., Sattelberger, A.-L., Sturmfels, B., and Telen, S. (2023). Bayesian integrals on toric varieties. *SIAM Journal on Applied Algebra and Geometry*, 7(1):77–103. arXiv:2204.06414.

Brown, W. M., Prager, E. M., Wang, A., and Wilson, A. C. (1982). Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J. Mol. Evol.*, 18:225–239.

Fevola, C., Mizera, S., and Telen, S. (2024a). Landau singularities revisited: Computational algebraic geometry for Feynman integrals. *Physical Review Letters*, 132:101601. arXiv:2311.14669.

Fevola, C., Mizera, S., and Telen, S. (2024b). Principal Landau determinants. *Comput. Phys. Commun.*, 303:109278. arXiv:2311.16219; package `PLD.jl`.

Fontaine, M. C. et al. (2015). Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347:1258524.

Table S15: The hominid tRNA-Phe quartet inside the hominoid quintet: exact log marginal likelihoods of the three splits of human (H), chimpanzee (C), gorilla (G) and orangutan (O) under JC69, and the margin $\Delta_{\mathrm{HG}} = \log Z(\mathrm{HG} \mid \mathrm{CO}) - \log Z(\mathrm{HC} \mid \mathrm{GO})$ (positive when the grouping of human with gorilla is preferred). Rows 1–2: quartet model, five exponential branches of rate 4/3. Rows 3–8: the prior induced by the five-taxon model with gibbon's row summed out (row 8, the equal-weight mixture of rows 3–7, in which the named branch is gamma of shape two and rate 4/3 and the other four exponential). Row 9: five-taxon model and data, log of the sum of Z over the five topologies containing each split. Row 10: the best single five-taxon topology containing each cherry (main-text Table 4). Rows 11–12: all five branches gamma of shape two. In every row $Z(\mathrm{HC} \mid \mathrm{GO}) = Z(\mathrm{HO} \mid \mathrm{CG})$ exactly except rows 3 and 5, which the human–gorilla exchange maps onto each other.

	model, prior, columns	$\log Z(\mathrm{HC} \mid \mathrm{GO})$	$\log Z(\mathrm{HG} \mid \mathrm{CO})$	$\log Z(\mathrm{HO} \mid \mathrm{CG})$	Δ_{HG}
1	quartet, Exp, 69 columns	−161.9379803	−161.9366609	−161.9379803	+0.0013194
2	quartet, Exp, 67 columns	−158.7179011	−158.7164889	−158.7179011	+0.0014122
3	gamma branch: human, 67	−161.9459985	−161.9392951	−161.9516040	+0.0067034
4	gamma branch: chimpanzee, 67	−161.2240800	−161.2283748	−161.2240800	−0.0042948
5	gamma branch: gorilla, 67	−161.9516040	−161.9392951	−161.9459985	+0.0123089
6	gamma branch: orangutan, 67	−160.7962330	−160.7962424	−160.7962330	−0.0000094
7	gamma branch: internal, 67	−162.5703963	−162.5660248	−162.5703963	+0.0043715
8	induced mixture of 3–7, 67	−161.5082824	−161.5066652	−161.5082824	+0.0016172
9	quintet, summed over gibbon's placement, 67	−189.3484577	−189.7002747	−189.3484577	−0.3518169
10	quintet, best single topology, 67	−189.6451165	−190.2076498	−189.6451165	−0.5625333
11	quartet, gamma shape 2 mean 0.75, 67	−167.1159821	−167.0953471	−167.1159821	+0.0206350
12	quartet, gamma shape 2 mean 0.1, 67	−151.5988442	−151.5833201	−151.5988442	+0.0155241

Gel'fand, I. M., Kapranov, M. M., and Zelevinsky, A. V. (1990). Generalized Euler integrals and A-hypergeometric functions. *Adv. Math.*, 84:255–271.

Goldman, N. (1993). Statistical tests of models of DNA substitution. *J. Mol. Evol.*, 36:182–198.

Hayasaka, K., Gojobori, T., and Horai, S. (1988). Molecular phylogeny and evolution of primate mitochondrial DNA. *Mol. Biol. Evol.*, 5:626–644.

Hendy, M. D. and Penny, D. (1989). A framework for the quantitative study of evolutionary trees. *Systematic Zoology*, 38(4):297–309.

Hofreiter, M., Jaenicke, V., Serre, D., von Haeseler, A., and Pääbo, S. (2001). DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Research*, 29:4793–4799.

Huh, J. (2013). The maximum likelihood degree of a very affine variety. *Compositio Mathematica*, 149:1245–1266. arXiv:1207.0553.

Jarvis, P. D., Bashford, J. D., and Sumner, J. G. (2005). Path integral formulation and Feynman rules for phylogenetic branching models. *Journal of Physics A: Mathematical and General*, 38(44):9621–9647. arXiv:q-bio/0411047.

Jensen, E. L., Quinzin, M. C., Miller, J. M., Russello, M. A., Garrick, R. C., Edwards, D. L., Glaberman, S., Chiari, Y., Poulakakis, N., Tapia, W., Gibbs, J. P., and Caccone, A. (2022). A

new lineage of Galapagos giant tortoises identified from museum samples. *Heredity*, 128:261–270.

Kontsevich, M. and Zagier, D. (2001). Periods. In *Mathematics Unlimited – 2001 and Beyond*, pages 771–808. Springer.

Krings, M., Stone, A., Schmitz, R. W., Krainitzki, H., Stoneking, M., and Pääbo, S. (1997). Neandertal DNA sequences and the origin of modern humans. *Cell*, 90:19–30.

Lartillot, N. and Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Syst. Biol.*, 55:195–207.

Lee, R. N. and Pomeransky, A. A. (2013). Critical points and number of master integrals. *JHEP*, 11:165. arXiv:1308.6676.

Messier, W. and Stewart, C.-B. (1997). Episodic adaptive evolution of primate lysozymes. *Nature*, 385:151–154.

Pachter, L. and Sturmfels, B. (2004). Tropical geometry of statistical models. *Proceedings of the National Academy of Sciences USA*, 101(46):16132–16137. arXiv:q-bio/0311009.

Rusakov, D. and Geiger, D. (2005). Asymptotic model selection for naive Bayesian networks. *Journal of Machine Learning Research*, 6:1–35.

Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., and Pääbo, S. (2012). Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS ONE*, 7:e34131.

Schwartz, M. D. (2026). JaCK & Jill: exact and certified Bayesian evidence for phylogenetic models. Manuscript in preparation.

Speyer, D. and Sturmfels, B. (2004). The tropical Grassmannian. *Advances in Geometry*, 4(3):389–411. arXiv:math/0304218.

Sturmfels, B. and Telen, S. (2021). Likelihood equations and scattering amplitudes. *Algebraic Statistics*, 12(2):167–186. arXiv:2012.05041.

Takayama, N., Kuriki, S., and Takemura, A. (2018). A-hypergeometric distributions and Newton polytopes. *Advances in Applied Mathematics*, 99:109–133. arXiv:1510.02269.

Thawornwattana, Y., Dalquen, D., and Yang, Z. (2018). Coalescent analysis of phylogenomic data confidently resolves the species relationships in the *Anopheles gambiae* species complex. *Mol. Biol. Evol.*, 35:2512–2527.

Turtle Taxonomy Working Group [A. G. J. Rhodin, J. B. Iverson, U. Fritz, N. Gallego-García, A. Georges, H. B. Shaffer, and P. P. van Dijk] (2025). Turtles of the world: Annotated checklist and atlas of taxonomy, synonymy, distribution, and conservation status (10th ed.). *Chelonian Research Monographs*, 10.

Wen, D., Yu, Y., Hahn, M. W., and Nakhleh, L. (2016). Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. *Molecular Ecology*, 25:2361–2372.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., and Chen, M.-H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.*, 60:150–160.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, 10:1396–1401.

Zeilberger, D. (1990). A holonomic systems approach to special functions identities. *J. Comput. Appl. Math.*, 32:321–368.