Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Decoding assumptions and inference about a robustness preference in codon reassignments along sampled histories

Jonathan Reich

jonathanreich100@gmail.com

Independent Researcher

Research Article

Keywords:

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11480587/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Decoding assumptions and inference about a robustness preference in codon reassignments along sampled histories

Jonathan Reich
Independent Researcher, Melbourne, Victoria, Australia
Corresponding author: jonathanreich100@gmail.com
ORCID: 0009-0000-0108-0717

Abstract

Catalogue-associated net transitions between genetic codes reach the silent third-position packing maximum more often than uniform targets do. Whether that signals selection depends on which targets decoding made accessible and which reassignments were detected. We prove that from a block-greedy code every single-box reassignment to an amino acid already in the codon's box keeps the code optimal, that a robustness coefficient in a conditional-logit target-choice model is identified exactly when the within-pool contrasts span it, and that the class-conditional null stays exact when the decoding state is uncertain. On 200 sampled net-transition histories, the excess over uniform targets persists (35.5 against 20.0 units per history), but a recording ratio of 1.85 for optimum-attaining units would bring its merged reference statistic above 0.05. Within accessibility classes no repertoire's test rejects once restrictive decoding states are treated as uncertain, and in conditional simulations pooled accessibility with no preference makes the fixed-state restrictive rule reject in all 60 replicates. In pointwise comparisons with fitted sequential generators, the observed statistic has upper-tail ranks below 0.05 for common preferences of odds 2, 2.5 and 3 under a restrictive repertoire (at most 0.025) and of odds 2.5 and 3 under pooled wobble rules, and above 0.05 at restrictive odds 1.5 (0.099) and pooled odds 2 (0.059); these ranks have no established exclusion coverage. Hierarchical intervals include zero under the primary priors. The data constrain several specified preference models but do not establish whether accessibility, a robustness preference or ascertainment generated the enrichment.

Keywords: genetic code evolution; codon reassignment; ancestral reconstruction; tRNA modification; identifiability; stochastic histories

1 Introduction

Variant genetic codes are often compared with the standard code on robustness to mutation and mistranslation. Błażej et al. (2019, p. 25) found that "the changes that make the code more robust evolve more often than the changes deteriorating its property", and cautioned that this "does not automatically imply that their reassignments were directly selected to minimize the mutational errors" (p. 30). This paper asks what can be inferred about such a preference once the reassignments are placed on reconstructed histories and the targets each reassignment could reach are modelled. The endpoint throughout is a structural one: whether the code produced by a reassignment attains the maximum number of silent third-position edges attainable with its own synonymy-class sizes and stop codons. We call it the *unweighted packing optimum*. It weights no amino-acid similarity, transition bias or codon usage, and it is not a translational error cost. For the standard code, the exact combinatorics of the unweighted silent third-position count, and of its counterpart restricted to transitions, are worked out in Reich (2026b).

A reassignment is a change in decoding, and the literature on its mechanisms has long held that this restricts the targets. Knight et al. (2001b, p. 299) found that codons which disappear from mitochondrial genomes are disproportionately reassigned, whereas codons predicted to disappear from composition are not; they tested the ambiguous-intermediate hypothesis against the share of all possible changes compatible with three mispairings (pp. 309–311), and read the reassignments that extend wobble pairing within a four-codon box as transient ambiguity followed by loss of the original tRNA (p. 311). Yokobori et al. (2001) described box-resident routes in detail: once tRNA-Arg is lost, tRNA-Ser(GCU) reads `AGA` without structural change, and m^7G34 extends it to `AGG` (p. 318); tRNA-Met can gain `AUA` through f^5C34 or t^6A37 with no change in its recognition by the synthetase (p. 320); and anticodon sequence, modified nucleotides and identity determinants "restrict the manner in which the code can change" (p. 323). Swire et al. (2005, p. 128) found that in sense-to-sense reassignments the moved codon was not rare in proxies of the ancestral genome, and argued that benefits from reduced mutational load are too weak to drive late reassignments (p. 134). Sengupta et al. (2007, pp. 662–663) placed the mechanisms in a gain–loss framework (codon disappearance, unassigned codon, ambiguous intermediate and compensatory change) and concluded that reassignments since the establishment of the standard code are chance events not guided by positive selection (pp. 684, 686). Sengupta and Higgs (2015, PDF p. 6) add that the small fitness differences between codes that translational error might create are not the drivers of code change in modern organisms, and that most sense-to-sense reassignments lie within one column of the code, needing only an anticodon change in one tRNA and loss of the original reader.

If the targets that decoding makes accessible are also those that attain the packing optimum, a uniform null attributes to target choice what is a property of decoding. The companion paper (Reich, 2026a) analyses this for catalogued origins scored from the standard code (table 1) with present-day decoders. The present paper asks the historical question. It places the same catalogue on sampled joint histories, so that every catalogue-associated net transition is scored on its reconstructed predecessor code under the specified grouping or order. It samples node states, rate categories and reader ancestry (with its replacement rate), uses fixed weights over topology alternatives, carries the uncertain decoding state into the tests, and assesses alternative branch lengths, observation models and layouts in sensitivity analyses.

What is shared and what is new. The event law and the rank condition (§§2–3) are restated in the form the histories need, and the identification corollary is the companion's. New here are: an exact null when the decoding state is uncertain, with a conservative bound (Proposition 10); histories with reader replacement, time-calibrated branch lengths and explicit order (§4); a benchmark of seven accessibility models against documented reader routes (§5); tests merged over histories and a hierarchical model with compartment and lineage slopes (§6); conditional simulations in which the history sampler itself is rerun on simulated data; and an exact conditional tilt identity for outcome-dependent ascertainment under a stipulated unit-level recording model. A reviewer archive with every analysed tree, the per-origin evidence and the test and fit tables accompanies the submission.

The main results are conditional. Against uniform targets the excess is large, but under a stipulated unit-level recording model outcome-dependent ascertainment makes its merged reference statistic exceed 0.05. Within accessibility classes none of the seven repertoires' tests rejects once restrictive decoding states are treated as uncertain. The excess that appears under the restrictive repertoire when its decoding state is fixed is generated, in conditional simulations that rerun the history sampler, by pooled accessibility with no preference, and in the hierarchical model it attenuates once neighbour status, co-occupancy and chemical similarity are adjusted for. Pointwise comparisons under fitted generators, in which each simulated unit chooses its target or target tuple on its simulated predecessor code, give the observed statistic upper-tail ranks below 0.05 for the tested common preferences of odds 2, 2.5 and 3 under a restrictive

repertoire and of odds 2.5 and 3 under pooled wobble rules, and ranks above 0.05 at restrictive odds 1.5 and pooled odds 2; they are ranks under fitted generators, not established exclusions. Hierarchical intervals include zero under the primary priors and, in the two-history calibration, had little power; equivalence to no preference at odds 1.5 is not demonstrated; and the benchmark against documented reader routes reveals differing failures but establishes no model as the ancestral repertoire.

2 The packing optimum and its event law

Setting. Let $\mathcal{C} = \{\mathrm{U}, \mathrm{C}, \mathrm{A}, \mathrm{G}\}^3$, and let the *box* of a prefix p be $\beta_p = \{p\mathrm{U}, p\mathrm{C}, p\mathrm{A}, p\mathrm{G}\}$. A code is a map f from $\mathcal{C}$ to amino acids and Stop (stop codons share one label). Write $n_\ell = |f^{-1}(\ell)|$ and $x_{\ell\beta} = |f^{-1}(\ell) \cap \beta|$; the *parts* of a sense class are its non-zero $x_{\ell\beta}$. The score is the number of silent third-position edges,

$$U_3(f) = \#\{\{x, y\} \subseteq \beta : \ x \neq y, \ f(x) = f(y)\},$$

the edges of the disjoint union of 16 cliques of size $q = 4$. Every result in this section holds for any partition of $\mathcal{C}$ into blocks of one size q. The *fixed-stop domain* $D_{\mathrm{fix}}(f)$ is the set of codes with the stop codons of f and the same multiset of sense-class sizes; f is *at its maximum* if $U_3(f) = \max_{D_{\mathrm{fix}}(f)} U_3$. A sense class is *block-greedy* if at most one of its parts differs from q, and a code is block-greedy if all its sense classes are. Let $\kappa(f)$ count the stop–stop edges and $\Gamma_q(n) = \lfloor n/q \rfloor \binom{q}{2} + \binom{n \bmod q}{2}$.

A reassignment changes the domain: the class sizes change, and so does the attainable maximum. "At its maximum" is therefore a statement about the code relative to its own domain, not about absolute robustness. A reassignment can lower U_3 and leave the code at its new maximum, and an ancestral code need not be at its maximum before the change.

Lemma 1 (Capacity-greedy bound)**.** *If $x_1, \ldots, x_r \in \{0, \ldots, q\}$ sum to n and γ is convex on $\{0, \ldots, q\}$ with $\gamma(0) = 0$, then $\sum_j \gamma(x_j) \leq \lfloor n/q \rfloor \gamma(q) + \gamma(n \bmod q)$.*

Proof. Extending x by a zero if necessary, let y have $\lfloor n/q \rfloor$ entries q, one entry $n \bmod q$ (omitted when it is zero) and zeros, padded to the length of x. The k largest entries of y sum to $\min(qk, n)$, and the k largest x_j sum to at most $\min(qk, n)$, since each is at most q and all sum to n. So y majorises x, and $x \mapsto \sum_j \gamma(x_j)$ is Schur-convex for convex γ (Marshall et al., 2011). □

Proposition 2 (Equality case)**.** *For every code g, $U_3(g) \leq G(g) := \kappa(g) + \sum_{\ell \text{ sense}} \Gamma_q(n_\ell)$, with equality if and only if g is block-greedy. Call g* tight *if $\max_{D_{\mathrm{fix}}(g)} U_3 = G(g)$. A block-greedy code is tight and at its maximum, and a tight code is at its maximum if and only if it is block-greedy.*

Proof. In $D_{\mathrm{fix}}(g)$ the stops are fixed, a stop–sense edge is never silent and a sense edge is silent if and only if both ends carry one label, so $U_3(g) = \kappa(g) + \sum_{\ell,\beta} \binom{x_{\ell\beta}}{2}$. Lemma 1 with $\gamma(x) = \binom{x}{2}$ bounds each class by $\Gamma_q(n_\ell)$. Parts in $\{1, \ldots, q\}$ with at most one part different from q form the greedy vector, which attains the bound. If two parts $a \geq b$ both lie in $\{1, \ldots, q-1\}$, moving one codon from the part b to the part a raises $\binom{a}{2} + \binom{b}{2}$ by $a - b + 1 > 0$, so the bound is not attained. (The exchange is an argument about the numerical bound of one class; it need not be realisable in a complete code.) A block-greedy g attains $G(g)$, which bounds the maximum, so it is tight and at its maximum; a tight code has maximum $G(g)$, attained if and only if it is block-greedy. □

Theorem 3 (Event law)**.** *Let f be block-greedy and let a change move $X \subseteq C = f^{-1}(s) \cap \beta$, $1 \leq k = |X| \leq c = |C|$, to a label $t \neq s$, with $a = x_{t\beta}(f)$. The new code is block-greedy if and only if* (S) *s is Stop, or $k = c$, or $c < q$, or $n_s \equiv 0 \pmod q$; and* (T) *t is Stop, or $a > 0$, or $k = q$, or $n_t \equiv 0 \pmod q$.*

Proof. Only the classes s and t change, and only in β; Stop imposes nothing. For s: its part c becomes $c-k$, which vanishes if $k=c$. If $c<q$, c was the unique non-q part of s, so s keeps at most one. If $c=q$ and $k<q$ a new non-q part appears, harmless if and only if s had no other, that is $n_s \equiv 0 \pmod q$. For t: if $a>0$ the box holds s and t, so $a \le q-c<q$ was the unique non-q part of t, and it becomes $a+k \le q$. If $a=0$, t gains a part k, harmless if and only if $k=q$ or $n_t \equiv 0 \pmod q$. □

Corollary 4 (Co-occupied changes)**.** *If f is block-greedy and the target already labels a codon of the box ($a>0$), the new code is block-greedy, hence tight and at its maximum.*

Proof. $a>0$ gives (T), and $c \le q-a<q$ gives (S); apply Proposition 2. □

Corollary 5 (One predicate for the strictly convex class)**.** *Let γ be strictly convex on $\{0,\dots,q\}$ with $\gamma(0)=0$, and $\Phi_\gamma(g)=\sum_{\ell \text{ sense},\beta}\gamma(x_{\ell\beta}(g))$. If $D_{\text{fix}}(g)$ contains a block-greedy code, g maximises Φ_γ over $D_{\text{fix}}(g)$ if and only if g is block-greedy, for every such γ at once.*

Proof. The exchange step of Proposition 2 uses only $\gamma(a+1)-\gamma(a)>\gamma(b)-\gamma(b-1)$ for $a \ge b$, which is strict convexity, so Φ_γ is bounded by its greedy value with equality if and only if g is block-greedy; a block-greedy member of $D_{\text{fix}}(g)$ attains the bound. □

Examples of γ are x^2 (a Simpson-type sum), x^p with $p>1$, and $x\log x$ (a negative-entropy sum), each strictly convex with $\gamma(0)=0$. On a domain that contains a block-greedy code these scores and U_3 have the same maximisers, so the at-maximum predicate is one predicate for the class. On a domain that contains none, the corollary says nothing, and the scores need not agree. The analysis below never relies on the equivalence: every outcome is scored against the U_3 maximum of its own domain, computed by a dynamic programme over the margins of D_{fix} (Reich, 2026a); its recurrence is stated in Online Resource 1 (Section **??**).

Theorem 6 (Joint changes)**.** *Let f be block-greedy and let changes $e_1,\dots,e_r$ move pairwise disjoint sets $X_i \subseteq f^{-1}(s_i)\cap\beta_i$ to labels t_i together. If every sense label occurs in $\{s_i,t_i\}$ for at most one i, the new code is block-greedy if and only if every e_i satisfies* (S) *and* (T) *in f.*

Proof. A sense class not among the s_i,t_i keeps its codons. A class among them is touched by exactly one e_i: its size changes only by e_i, and its part in β_i changes only by e_i, since another change acting in β_i moves codons of other classes. So it is block-greedy after all the changes exactly when it is after e_i alone. □

The hypothesis cannot be dropped, which is why joint episodes are scored on the joint code rather than by combining their members' outcomes. Corollary 4 is the arithmetic of the box-resident route: when a reader already in the box extends its reading, the moved codons go to a co-occupant of the box, and from a block-greedy code that outcome attains the optimum whatever its probability of arising. Table 1 is block-greedy, so the event law decides which single changes from it yield block-greedy codes. Reconstructed ancestors are another matter: of the 79 distinct ancestral codes from which the catalogue changes of the sampled histories (§4) are made, 27 are block-greedy and 27 are at their maximum, and a share 0.66 of the distinct transitions start from a block-greedy code. That is why outcomes are computed rather than read from the law.

3 What target choices can identify

Let a target-choice model give candidate j of unit i the weight $m_j \exp(\theta\cdot z_{ij})$, where $m_j>0$ does not depend on θ and $z_{ij}=(x_{ij},r_{ij})$ stacks accessibility features x and the robustness outcome r; the robustness coefficient β is the last component of θ. Let $\mathcal{D}$ be the set of within-pool differences $z_{ij}-z_{ik}$ over all units and pairs of candidates.

Proposition 7 (Rank condition)**.** *With the pools, the multiplicities m and the features z fixed, and θ ranging over all of $\mathbb{R}^p$, two parameter values θ, θ' give the same distribution of choices if and only if $\theta - \theta'$ is orthogonal to $\mathcal{D}$. Hence a linear combination $a \cdot \theta$ is identified if and only if a lies in the span of $\mathcal{D}$; in particular β is identified if and only if the unit vector of r lies in that span.*

Proof. The probability of j in unit i is $m_j / \sum_k m_k \exp\{\theta \cdot (z_{ik} - z_{ij})\}$, which depends on θ only through the inner products $\theta \cdot d$, $d \in \mathcal{D}$. If $\theta - \theta' \perp \mathcal{D}$ these agree, and so do all choice probabilities. Conversely, if $(\theta - \theta') \cdot d \neq 0$ for some $d = z_{ij} - z_{ik}$, the odds of j against k in unit i differ. So $a \cdot \theta$ is constant on every equivalence class $\theta + \mathcal{D}^{\perp}$ exactly when $a \in (\mathcal{D}^{\perp})^{\perp} = \operatorname{span} \mathcal{D}$. □

Corollary 8 (Co-occupancy and the optimum)**.** *Let every unit be a single change from a block-greedy code, with features c (co-occupancy) and y (at maximum) and coefficients (γ, β), any further weight being absorbed into m. Candidates are of type $A = (1, 1)$, $B = (0, 1)$ or $C = (0, 0)$; $(1, 0)$ cannot occur (Corollary 4). Pairs A–C, A–B and B–C contribute the contrasts $(1, 1)$, $(1, 0)$ and $(0, 1)$. (γ, β) is identified if and only if at least two contrast types occur. If no pool holds a type-B candidate together with one of another type, the choices depend on (γ, β) only through $\gamma + \beta$, which is identified if some pool holds an A–C contrast and otherwise is not.*

Proof. Two of the three contrast vectors span $\mathbb{R}^2$, one does not; apply Proposition 7. Without a type-B pair only $(1, 1)$ can occur, and it identifies $\gamma + \beta$ when it does. □

This is the identification statement of the companion paper (Reich, 2026a), and the two papers use it identically. It is structural: it holds for every accessibility weighting. What is not structural is which contrasts the data contain. A uniform comparison stratifies neither by accessibility nor by co-occupancy, so its excess mixes the co-occupancy and optimum preferences with every accessibility effect absorbed into m, and cannot by itself be attributed to β.

Proposition 9 (Class-conditional null)**.** *If, within a unit, the weight of an (uncompressed) candidate depends on it only through a class $\kappa(j)$, then given the class of the observed choice (an event of positive probability) the choice is uniform among the candidates of that class, whatever the class weights.*

Proof. Every candidate of class κ has weight ω_κ, so conditional on the class the probabilities are equal. □

When identical candidates are compressed into one row, the row's probability is proportional to its multiplicity. The class-conditional null of §6 is exact under this stated model: the history's pools and classes held fixed, class-only accessibility weights, and units independent. It is not exact for the whole procedure in which the histories are themselves inferred from the same targets; that is examined by simulation (§6).

The classes of a restrictive repertoire depend on readings, such as modified wobble bases, whose availability at the ancestral node is uncertain. Let unit i have a decoding state σ_i in a finite support Σ_i^+, with any joint law across units, and let state s give candidate j the class $c_s(j)$. The *cell* of j is $K_i(j) = \{k : \ c_s(k) = c_s(j) \text{ for every } s \in \Sigma_i^+\}$, the common refinement of the states' partitions.

Proposition 10 (Uncertain decoding states)**.** *Suppose that under no preference, given the states, units choose independently with weights that depend on a candidate only through its class in the unit's state. Then, conditional on the cells of the observed choices (events of positive probability), the choices are independent and uniform on their cells, whatever the joint law of the states and the class weights. If p_i^+ is the largest over $s \in \Sigma_i^+$ of the share of optimum-attaining candidates in the observed choice's class in state s, the upper tail of the Poisson-binomial law with probabilities p_i^+ is a valid P-value.*

Proof. Fix states σ in the supports. A cell lies within one class of every state of the support, so a candidate's weight is constant on its cell, and given σ and the cells the choices are independent and uniform on their cells. That law does not involve σ, so averaging over the conditional law of σ leaves it unchanged. For the bound, the upper tail of a sum of independent Bernoulli variables is non-decreasing in each probability, so it is at least the fixed-state P at the true state, which is valid conditionally on σ and therefore unconditionally. □

Under a common coefficient θ the same argument gives the exponential tilt of the cell law, so the tilt interval is exact for θ. Weighting the state-wise class-conditional laws by the prior state probabilities, as the earlier version of this analysis did, is not exact in general: the observed class updates the state probabilities through the unknown class weights. With 100 candidates, two of them optimal, and two equally likely states that swap which optimum sits in an easy class of 10, a level-0.06 test built that way rejects with probability about 0.1. The cell null conditions on more, so fewer units carry information. It is primary below, the bound is reported beside it, and the prior mixture is kept for comparison only. The prior enters the cell null only through its support, and the argument holds for states shared across units through common enzymes as well as for independent ones.

What the data contain. Corollary 8 describes only part of the design. Joint episodes and non-block-greedy ancestors can hold a co-occupied candidate that fails the optimum: 6–11 units per sampled history contain one. The general condition of Proposition 7 was therefore checked directly on the reach and co-occupancy features (§5) in every history: the robustness coefficient is identified in all 200 histories under `main`, `compr` and `compri` (the last two at a fixed decoding state). Per history, 45–62 units have a pool that varies in the optimum, the only units that can inform β; of these, 43–60 hold two candidates with the same reach and co-occupancy features but different outcomes, and 0–3 vary only together with those features. Adjusting for co-occupancy therefore leaves β identified, by a few dozen units. The difficulty is imprecision and dependence on assumptions, not structural non-identification.

4 Sampled histories

Catalogue and trees. The catalogue has 83 lineage-level reassignment rows with core evidence (confirmed, supported or computational), phages excluded. They were placed on six trees: mitochondria, plastids, bacteria, nuclear eukaryotes, ciliates and Saccharomycotina yeasts. These are induced subtrees of the Open Tree of Life synthesis, with the Sphaeropleales from published chloroplast phylogenomic trees (Fučíková et al., 2016) and the ciliate and yeast trees of Kollmar and Mühlhausen (2017, Figs 1–2). Branch lengths are divergence times from TimeTree (Kumar et al., 2022): each internal node is dated by the median of pairwise ages between its children's tips, which dates 631 of the 919 internal nodes of the analysed primary trees, and an undated node is interpolated on rank-based heights between its dated neighbours; ages are then made monotone along every path and very short edges are floored at a small fraction of the root age. Before that repair at most 3 parent nodes per tree are younger than a child, each a conflict between two dated nodes, and after it there are 0. Rank-based lengths (Grafen, 1989) and unit lengths are sensitivity samplers. Time is a proxy for the opportunity to change, not a rate of code change.

Characters and process. Codons that a catalogue row moves together form one character, whose states are tuples of meanings; a tip that knows only some of the codons is compatible with every tuple that agrees on them, and a codon absent from a genome is unknown. Characters are the connected components of the codon sets of the catalogue rows carried on a tree, so they do not overlap. A character's states are the tuples observed complete at some tip, the root

tuple, and every completion of a partially observed tip in which each unknown codon takes a meaning that codon shows elsewhere on the tree; no other combination is admitted. Of the 25 characters on the primary trees, 7 have more than one codon; the largest state space has 23 tuples (mitochondrial `AGA`+`AGG`) and the others have at most 6. Each character evolves by a symmetric Mk process: a state is left at the character's rate r, towards each of the other $k-1$ states at rate $r/(k-1)$, so a particular transition is rarer in a larger state space. One jump may change several codons, to one meaning or to different ones. A process restricted to one codon per jump is not defined on these state sets, because several multi-codon characters would split into mutually unreachable states; on the full product of meanings with a generator that factorises over codons it is a set of per-codon processes, independent given their shared rate, which is the unlinked layout below. Rates across characters are drawn from a discretised gamma distribution (eight categories; Yang, 1994) whose shape and mean are fitted for each tree and topology; every draw samples each character's rate category, so uncertainty in the category is carried through, but uncertainty in the fitted shape and mean is not. The root carries the table-1 meanings (a flat root is a sensitivity). Joint node states are drawn exactly from their posterior by a pruning pass and forward sampling, 4,800 draws per tree, and 200 joint histories are kept, each with weight 1/200. A branch whose end states differ carries one recorded change, so the histories are net changes between sampled nodes, not complete paths: reversals and repeated changes within a branch are not recorded, and each history reports the expected number of such hidden changes under its rates (11.5 per history, 9.6–13.9, against 86.3 recorded catalogue changes). On a branch, the codons of a character whose meaning changed are grouped by their ancestral and new meanings, and each group is one elementary change, moving codons of one label to one new label; codons that a row moves together and that make the same move on one branch count as one change. Every elementary change lies within one box (checked for every sampled change), and applying a branch's changes to its parent code gives the sampled code at its child. Probabilities over histories that are tabulated from all 4,800 draws per tree (origin counts and joint groups) have Monte Carlo standard error at most 0.007 (treating the exact posterior draws as independent); quantities computed on the 200 retained histories are less precise.

Tip states and their error. Each tip–codon meaning comes from the first available of: a state read from a source figure or text, a source-stated state of the tip table, the target of a catalogue row the tip carries, and the NCBI genetic-code assignment. Of the 2,011 cells, 72 come from figures read here, 233 from sources, 201 from catalogue carriers, 309 from NCBI codes set at genus rank or below and 1,145 from NCBI codes inherited from a higher rank; the remaining 51 are cells whose literature or source entry is absent or unresolved, which are unknown and enter the likelihood as missing. Literature, source and carrier states are treated as error-free. An NCBI state is wrong with probability 0.05 (set at genus rank or below) or 0.10 (inherited), the error spread evenly over the other states; exact tips, doubled error and the earlier reconstruction's tip rule are sensitivity samplers. Two further samplers relax the independence of these errors. In the first, tips that inherit an NCBI default from the same higher-rank taxon are wrong together, with the same marginal probability; this is the fully coupled extreme, and it overturns some large shared defaults (mitochondrial `UGA`→Trp merges rise from 0.07–0.31 to 0.71–0.75, and catalogue changes per history fall to 81.8). In the second, catalogue carriers whose row rests on computational evidence are wrong with probability 0.05 instead of being exact, so a single deviating tip can be an error rather than a change (78.3 catalogue changes per history). Both are carried through the exact tests and the hierarchical fits (Table **??**); the fixed-state restrictive excess persists under both, and the global coefficients change little.

Reader ancestry. Codon states cannot see tRNA ancestry. Two meanings have sourced reader ancestry. The `AGG`→Ala tRNA(CCU) of the Sphaeropleales derives from a duplicated tRNA-Ala(UGC) in *Chlorotetraedron* and *Neochloris* and from a remodelled tRNA in the Scenedes-

maceae (Žihala and Eliáš, 2019, pp. 2999–3000; Noutahi et al., 2019, Table 2, p. 773), and the CUG→Ser tRNAs of the Ser1 and Ser2 yeasts have different sources (Krassowski et al., 2018, PDF p. 4). Different reader ancestry does not by itself require different origins of the meaning: the reader can be replaced while the meaning persists. Each origin of a meaning therefore recruits a reader from one of $C = 5$ classes, and while the meaning persists the reader is replaced at rate λ, drawing a new class; λ has a log-uniform prior from 0.01 to 30 replacements per root-to-tip depth and is sampled jointly with the history by exact rejection. Readers of different ancestry can share an origin, at a cost that grows with the replacements it needs. A derived reader still orders its change: the `UAG`→Ala tRNA(CUA) derives from the new tRNA-Ala(CCU) (Žihala and Eliáš, 2019, p. 2999), so `AGG`→Ala precedes it. In the mitochondrial tree the posterior favours fast replacement ($P(\lambda \geq 1) = 0.84$); the codon states favour one Sphaeropleales origin, and the `AGG`→Ala lineages of *Chlorotetraedron* and the Scenedesmaceae share an origin with probability 0.90 (1.85 origins on average). The yeast rate stays slow ($P(\lambda \leq 0.1) = 0.81$; 2.10 CUG→Ser origins). The earlier rule that readers of different ancestry cannot share an origin is the $\lambda \to 0$ limit of this model; as a sensitivity sampler it gives 3.02 and 2.49 origins and merge probability 0.00, and dropping the reader data gives 1.93, 2.38 and 0.87. The reader data therefore discriminate little for this merge (0.90 with them, 0.87 without): the replacement model accommodates the discordant readers, and fast turnover is the way it reconciles them with the codon states, not an independent biological finding. Restricting the prior on λ moves the merge probability from 0.78 (rates at most 1) and 0.58 (at most 0.3) to 0.96 (at least 1). A recruitment class is an exchangeable label for an independent recruitment event; the classes are equally likely a priori, $C = 3$ and $C = 20$ are sensitivities, and λ is informed by few tips.

Topology and literature states. Contested clades enter as factors whose levels carry equal, fixed weights: three Sphaeropleales trees, two placements each of the Xenacoelomorpha, the Amoebozoa, the Mesodiniea, the CUG-Ser2 yeasts and the Ulvophyceae, and two placements of the GTDB placeholder genomes, apart or as one clade (Shulgina and Eddy, 2021, PDF pp. 10, 12). States read from Žihala and Eliáš (2019, Fig. 2, p. 2997) replace NCBI submitter defaults for Sphaeropleales tips, and Barcytė et al. (2025, Fig. 8) give the pedinophyte codes and place `UGA`→Trp of *Oistococcus* on its terminal branch after the Marsupiomonadales changes. A sampler without these states is a sensitivity. The topology weights are not updated by the likelihood, so the mixture over topologies is a stated modular weighting, not a topology posterior; whether the modal origin count depends on the topology is reported for every contested group (Table 1).

From histories to transitions. A *catalogue change* is the most recent change of a moved codon on the path to a carrier of a catalogue row that still has the row's target; reversals to the table-1 meaning and other changes are classified apart. The primary design takes the catalogue changes of every history: 547 distinct transitions, 74–94 per history in 64–84 target-choice units. Every sampled change (1193 distinct transitions, 83–100 units per history) is a sensitivity design. Each change is scored on the code at its branch's parent node, with the sampled states of every character and any change ordered before it on the branch. Per history, 32.3 (29–36) of 86.3 catalogue changes start from a code that differs from table 1, and 12.0 (7–18) reversals occur.

Origin counts and episodes. Origin counts are read from the histories as the number of distinct changes realising a group's rows (Table 1); the modal count is the same under every topology for 14 of the 14 contested groups. Changes of different characters on one branch, unordered by a reader rule, form one unresolved branch group, with 8.0 (6–10) per history and 10.1 branches per history carrying more than one change. Table 2 gives the probability that the members of each candidate group share a branch. Two changes on one branch need not be one molecular event. The primary design scores a group as one draw from a product pool on the common ancestral code (§6); a second layout writes every history as two order replicates of

Group	Tree	Mean	Mode (P)	Catalogue, conservative	liberal
M25	mito	1.01	1 (0.99)	–	–
MAUA_chloro	mito	1.99	2 (0.99)	–	–
MAUA_met	mito	1.17	1 (0.87)	1	3
MAUA_yeast	mito	1.00	1 (1.00)	1	2
MCGG_leu	mito	1.97	2 (0.97)	1	2
MUCG_stop	mito	1.03	1 (0.97)	–	–
MUGA_opisth	mito	3.30	3 (0.57)	1	5
MUGA_stram	mito	1.72	2 (0.66)	1	2
NUAR_cil	cil	6.94	7 (0.57)	1	6
NUAR_ulvo	nuc	1.47	1 (0.76)	1	3
NUGA_cil	cil	4.07	5 (0.40)	1	3
NUGA_cys	cil	2.04	2 (0.96)	1	2
set:sphaeropleales_AGG_Ala	mito	1.85	2 (0.63)	–	–
set:yeast_CUG_Ser	yeast	2.10	2 (0.89)	–	–

Table 1: Origin counts of the contested catalogue groups over the sampled histories, as modular posterior summaries under fixed topology weights and fitted rates (all 4,800 draws per tree): mean, modal count and its probability, with the catalogue's conservative and liberal counts where the group is listed there.

Lineage	Members	P(one branch)	Range over topologies
Saccharomyces mt	`AUA`→Met; `CUN`→Thr	1.00	0.99–1.00
Marsupiomonadales mt	`AGR`→Ala; `UUR`→Stop	0.94	0.92–0.96
Blastocrithidia nuclear	`UAR`→Glu; `UGA`→Trp	0.99	0.99–0.99
peDinoflagellate plastid	`AUA`→Met; `UUA`→Stop; `UCA`→Stop	1.00	1.00–1.00
bilaterian mt	`AUA`→Met; `AGR`→Ser	0.41	0.38–0.45
Pycnococcus mt	`UGA`→Trp; `AUA`→Met	0.49	0.46–0.54
Condylostoma nuclear	`UAR`→Gln; `UGA`→Trp	0.62	0.62–0.63
Chromochloris mt	`AGG`→Met; `CGG`→Leu	0.76	0.34–0.98

Table 2: Candidate joint episodes: the probability, over the sampled histories, that changes realising each member sit on one branch unordered. These are the episodes that the earlier node-marginal reconstruction treated as joint; the probability is now read from joint histories.

weight $1/(2 \times 200)$, in which the changes of a branch are sequential, each in an order drawn with its model probability and each scored on its predecessor's code. In a third, the unlinked sampler, every codon is its own character, so moves that a catalogue row makes together are not tied. All three are analysed (Table **??**). The product pool is a reference law for an unresolved order, not a molecular event model; the ordered layout, which scores sequential net transitions, gives the same repertoire dependence and the same broad global intervals (Table **??**); the simulation comparisons were not run on it.

Six merge histories. The earlier reconstruction enumerated 6 merge histories from node marginals (for example a Dikarya `UGA`→Trp merge with marginal 0.47). Node marginals are not probabilities of origin histories, and that reconstruction is superseded for inference. Its histories are kept only for analyses that need many cheap refits (the observation-flow and unit-set comparisons, the co-occupancy adjustment, the graded measures, the prior grid and a restricted calibration), and every number that refers to them is labelled.

5 Decoding accessibility

Reach. Every candidate target t of a transition moving X in box β of code f gets a reach class on f: the least number of anticodon substitutions (plus one for a change of charging identity, where a model allows it) that turn a reader in the repertoire into a reader of t whose base-34 reading ρ, under the model's rules, reads every codon of X and otherwise only codons already assigned to t ($X \subseteq \beta + \rho \subseteq X \cup f^{-1}(t)$). With no substitution the reader must keep its own codons. Classes are 0, 1, at least 2, none, and stop. This admits the box-resident route, in which a Met reader that gains `AUA` keeps reading `AUG`, and excludes readers that would also read codons keeping another meaning. Reach is an assumed accessibility proxy, as in the companion paper: it counts recognition changes and is not a probability of reassignment.

Seven models. `main` uses every reading of Knight et al. (2001a, Table 1) with the minimal reference repertoire on the ancestral code; `comp` the compartment's readings and repertoire; `compr` as `comp` without readings known only from variant codes (f^5C, m^7G34, $G\Psi$, edited U); `pooli`, `compi` and `compri` as `main`, `comp` and `compr` with one extra substitution for an identity change; and `lineage` the compartment's readings read by the tRNA genes of the sister genome. Where an ancestral box-class is one the minimal repertoire does not list (Ser on `UCU`, `UCC`, `UCG` once `UCA` is a stop), its existing readers are the fewest pooled readings that read exactly its codons, ties broken in a fixed order. That rule describes readers the ancestral code must have had and their present codons; the reading of any new route still comes from the model's own rules. The pooled rules credit every tRNA with readings found in any organism; the restricted repertoire withholds those documented only in variant lineages. Neither is an ancestral reconstruction, and the historical process need not lie between them.

Decoding states. For the plastid `AUA`→Met changes of the peDinoflagellate plastid and of *Chloroparvula*, tilS and trnI(cau) are absent on the ancestral side (Barcytė et al., 2025, Fig. 8B; Turmel et al., 2019, p. 1280), so lysidine is withheld there. For the bilaterian `AUA`→Met, ALKBH1, which oxidises m^5C34 to f^5C34 (Haag et al., 2016; Nakano et al., 2016), is present in `AUA`-Ile placozoan and cnidarian outgroups (§7).

Benchmark against documented routes. The seven models were scored, without change, against 56 documented reader origins (54 events, 46 conservative origin groups, the primary unit) and against the observed targets of 69 transitions (59 groups), beside a decoding-cost model and three baselines (Table 3). Every reach model names the documented route far more often than the uniform, neighbour-only and co-occupancy-only baselines: hit rates 0.38–0.47 against 0.008, 0.050 and 0.078, and the paired gain over co-occupancy (0.30–0.40) has a 95% lower bound of at least 0.18 (Table **??**). The other metrics are less favourable, and they matter for the model that produces the restrictive excess. Route coverage, the share of groups whose documented route a model can reach at all, is 0.52–0.85; on mean normalised rank `compr` does worse than neighbour-only (paired difference 0.106 (0.004 to 0.205)). On the 17 groups whose route is directly documented, the hit rate is 0.68 for `main` and 0.30 for `compri`. On the target task the paired hit-rate differences from co-occupancy-only range from −0.098 (−0.191 to −0.004) for `compri` to 0.09, and `compri` makes a share 0.67 of the other candidates as accessible as the observed target. Reaching the observed amino acid by some route is not the same as reaching it by the documented one (the paired and direct-evidence results are in Table **??**). The marginal hit-rate intervals of the seven models overlap, which is not a paired comparison and not evidence that they are equivalent. The seven models were fixed before the benchmark, but their rules and the documented routes draw on overlapping literature, so the benchmark is not independent validation. It supports none of the models as the historical repertoire; they are

	Documented routes (46 groups)		Observed targets (59 groups)	
Model	Hit rate (95% CI)	Coverage	Hit rate (95% CI)	False accessibility
`main`	0.47 (0.33–0.61)	0.68	0.36 (0.27–0.46)	0.31
`comp`	0.43 (0.28–0.57)	0.64	0.40 (0.29–0.50)	0.29
`compr`	0.47 (0.33–0.62)	0.52	0.30 (0.20–0.41)	0.29
`pooli`	0.45 (0.32–0.58)	0.85	0.36 (0.26–0.46)	0.44
`compi`	0.38 (0.25–0.50)	0.80	0.36 (0.26–0.46)	0.45
`compri`	0.42 (0.29–0.55)	0.68	0.22 (0.14–0.31)	0.67
`lineage`	0.45 (0.30–0.59)	0.66	0.40 (0.30–0.51)	0.28
cost model, compartment	0.50 (0.36–0.63)	0.87	0.23 (0.16–0.33)	0.26
cost model, full	0.48 (0.35–0.61)	0.93	0.24 (0.15–0.33)	0.27
uniform	0.008 (0.007–0.008)	1.00	0.051 (0.051–0.052)	1.00
neighbour-only	0.050 (0.033–0.068)	1.00	0.14 (0.12–0.15)	0.40
co-occupancy-only	0.078 (0.055–0.10)	1.00	0.32 (0.22–0.41)	0.57

Table 3: Accessibility models against documented reader histories. Hit: the probability that the documented route (or observed target) is ranked first, ties broken at random, averaged over conservative origin groups; intervals from a bootstrap over groups. Coverage: share of groups whose documented route the model can reach at all. False accessibility: share of the other candidate targets at least as accessible as the observed one. The cost model's route scores are in sample (its costs were set after the documented routes were read); the seven reach models were fixed before this benchmark, but their rules and the documented routes draw on overlapping literature.

imperfect accessibility proxies with different failure modes, and all seven are carried through the analysis.

The cost model. The decoding-cost model prices a route additively, in units of one anticodon substitution, with rate $\exp(-C)$. Components and defaults: substitution 1; duplication 1; charging-identity change 1 for Ala and Leu, whose synthetases do not use the anticodon (Yokobori et al., 2001, p. 322; Sengupta et al., 2007, p. 685), and by assumption for Ser, and 2 otherwise; base-34 reading 0 for a Crick reading, 0.5 for a generic reading listed for the compartment, 1.5 for one listed only for other compartments and 2 for a variant-only reading, plus 1 when a reading documented on one tRNA species is moved to another; C34 editing 2; near-cognate reading 2 or 3 by mismatch position; 1.5 for each other class the new reader would also read; and 0.5 when the old reader survives. It reaches recharging, editing and near-cognate routes the reach models cannot (coverage 0.87–0.93 against 0.52–0.85), but its route hit rate is no better than theirs, it stays between 0.39 and 0.54 over its one-at-a-time ranges, and its lead is in sample. It is reported as a benchmark, not used as an accessibility model.

Modification-availability priors. For `compr` and `compri` the readings whose availability would change a candidate's class are uncertain. The support of each reading's state, and a prior probability that the enzyme activity it needs was present at the ancestral node, are: 0.5 (0.3–0.8) for f^5C at the bilaterian `AUA`→Met, where ALKBH1 predates the change; 0.1 (0.05–0.3) for *Chloroparvula*, where the modified tRNA-Met emerged after the loss of trnI(cau) and tilS (Turmel et al., 2019, p. 1290); 0.2 (0.05–0.5) for the yeast `AUA`→Met and 0.2 (0.05–0.5) for m^7G34 at `AGR`→Ser, both of which follow a tRNA loss (Sengupta et al., 2007, pp. 673–675); a generic 0.3 (0.1–0.6) for other box-resident origins; and 0.1 (0–0.25) for any other relevant reading; the lower end of the last range excludes those readings altogether. The generic values are guesses and are flagged as such. 330 of the 547 sampled transitions have relevant readings. Classes are recomputed in every reading state. The exact tests use the cell null of Proposition 10, which depends on the priors only through which states have positive probability, with the worst-state

Model	Obs.	Unif. E	Unif. P	Unif. largest P	Class E	Class P	$\bar{P}$ (range)	Inform.
`main`	35.5	20.0	1.1×10^{-4}	0.0019	35.7	1.00	0.67 (0.51–0.86)	8–12
`comp`	35.5	20.0	1.1×10^{-4}	0.0019	36.9	1.00	0.81 (0.59–0.93)	19–29
`compr`	35.5	20.0	1.1×10^{-4}	0.0019	35.8	1.00	0.70 (0.55–0.88)	8–12
`pooli`	35.5	20.0	1.1×10^{-4}	0.0019	35.9	1.00	0.72 (0.59–0.87)	9–13
`compi`	35.5	20.0	1.1×10^{-4}	0.0019	36.8	1.00	0.79 (0.61–0.93)	20–31
`compri`	35.5	20.0	1.1×10^{-4}	0.0019	36.0	1.00	0.74 (0.61–0.89)	9–13
`lineage`	35.5	20.0	1.1×10^{-4}	0.0019	36.9	1.00	0.80 (0.59–0.93)	19–29

Table 4: Units attaining the fixed-stop U_3 maximum on their ancestral code, over the 200 sampled histories (catalogue design; `compr` and `compri` with the default modification-availability priors). Observed and expected: means over histories. Uniform P: merged, $\min(1, 2\bar{P})$; largest P: the largest per-history upper-tail P. Class P: the merged class-conditional P, $\min(1, 2\bar{P})$; `compr` and `compri` under the cell null. $\bar{P}$: the mean per-history P with its range, descriptive only. Inform.: range over histories of the units whose class stratum holds both outcomes.

bound and the earlier prior mixture beside it; the hierarchical model integrates the state by a latent-state mixture likelihood, which uses the prior values and assumes independent states. The fixed state (every relevant reading withheld) and the low and high priors are reported beside the default.

6 Target choice along the histories

Units and the episode law. A unit is a transition or a joint episode. A joint episode is one draw from the product of its members' pools, scored by whether the joint code attains its maximum: one indicator per episode. Each member's reach is computed on the common ancestral code, and sequential changes are separate units, each on its predecessor's code. The statistic counts units that attain the optimum. On table 1, with the companion's episodes, this law gives the same exact null distribution of the joint-code indicator count as the companion package's independent implementation, for both groupings of episodes and for both the uniform and the class-conditional draw (branch grouping, class-conditional $P = 0.58$).

Observation set. Every unit of a history enters every analysis under every model, so the unit and random-effect index sets are identical across models (Table **??**). No observed target is inaccessible under any model (0 cases). A model changes only which units carry information: a unit whose class stratum is constant in the optimum adds a constant to the statistic and its null, and a unit whose candidates all share features and outcome has a constant likelihood. On the six merge histories, keeping those constant units, dropping them per model, or restricting to the units informative under every model (56–58 units) changes the coefficients only in the second decimal and leaves the merged P unchanged (Table **??**).

Merging over histories. Each per-history P is exact under the stated model with that history taken as true. Two merged values are reported: $\min(1, 2\bar{P})$, where $\bar{P}$ is the mean per-history P; and the largest per-history P. The first is valid under any dependence when every component is valid under the same null distribution, and the second when the true history is one of those tested. Neither condition holds automatically here: a test valid when its own history is true need not be valid when another history generated the data, and 200 sampled node-resolution histories need not include the true one. Both merged values are therefore treated as statistics whose size is established by simulation of the whole procedure (§6). The mean $\bar{P}$ itself is descriptive and is not a P-value.

Model	Priors	Null	Class E	P $(2\bar{P})$	Largest P	$\bar{P}$	Inform.
`compr`	–	fixed state	35.8	1.00	0.80	0.63	22–33
`compr`	default	cell (exact)	35.8	1.00	0.88	0.70	8–12
`compr`	default	worst state (bound)	38.7	1.00	0.99	0.97	17–26
`compr`	default	prior mixture (round 2)	35.3	1.00	0.72	0.55	35–52
`compr`	low	cell (exact)	37.1	1.00	0.94	0.83	19–29
`compr`	low	worst state (bound)	37.1	1.00	0.94	0.83	19–29
`compr`	low	prior mixture (round 2)	35.9	1.00	0.82	0.65	22–33
`compr`	high	cell (exact)	35.8	1.00	0.88	0.70	8–12
`compr`	high	worst state (bound)	38.7	1.00	0.99	0.97	17–26
`compr`	high	prior mixture (round 2)	35.2	1.00	0.70	0.54	35–52
`compri`	–	fixed state	24.9	0.0041	0.035	0.0021	37–53
`compri`	default	cell (exact)	36.0	1.00	0.89	0.74	9–13
`compri`	default	worst state (bound)	38.7	1.00	0.99	0.97	19–28
`compri`	default	prior mixture (round 2)	28.6	0.052	0.17	0.026	38–54
`compri`	low	cell (exact)	30.9	0.20	0.68	0.100	28–43
`compri`	low	worst state (bound)	30.9	0.20	0.68	0.100	28–43
`compri`	low	prior mixture (round 2)	25.5	0.0073	0.053	0.0036	37–53
`compri`	high	cell (exact)	36.0	1.00	0.89	0.74	9–13
`compri`	high	worst state (bound)	38.7	1.00	0.99	0.97	19–28
`compri`	high	prior mixture (round 2)	32.3	0.35	0.45	0.18	38–54

Table 5: The restrictive models under the decoding-state constructions, over the 200 sampled histories: the fixed state; the exact cell null (Proposition 10); the worst-state bound; and the earlier prior mixture, which is not exact. Inform.: range over histories of the units whose conditioning stratum holds both outcomes.

Uniform and class-conditional tests. Against uniform targets, 35.5 units per history attain the optimum against 20.0 expected; merged $P = 1.1 \times 10^{-4}$, and the largest per-history P is 0.0019 (Table 4). Within reach classes no excess is detected under the pooled, compartment and lineage repertoires: under `main` 35.7 are expected, merged $P = 1.00$, with 8–12 informative units per history. The restrictive repertoires need a decoding state (Table 5). At a fixed state, with every relevant reading withheld, `compri` expects 24.9, merged $P = 0.0041$, and every history falls below 0.05 (largest $P = 0.035$); `compr` does not ($P = 1.00$). With the state uncertain under the default priors, the exact cell null gives merged $P = 1.00$ for `compri`, the worst-state bound 1.00, and the earlier prior mixture, which is not exact, 0.052. The cell null is not evidence that the excess is absent. It keeps 9–13 informative units per history against 37–53 at the fixed state, because a unit informs it only when the cell of its observed choice holds both outcomes. The cell null depends on the priors only through their support: the high priors, which change probabilities but no support, give the same cells as the default. The low priors set the probability of any other relevant reading to zero, which removes those states from the support, refines fewer cells, and gives 0.20 on 28–43 informative units. A small positive probability and an impossible state therefore have very different consequences for this test. The restrictive excess is therefore a statement about a known decoding state, and with the state uncertain the primary uncertain-state cell rule does not reject on these data.

The hierarchical robustness coefficient. A conditional-logit model gives candidate j of unit i the weight $m_j \exp(\alpha \cdot x_{ij} + b_i r_{ij})$, with reach indicators x ($d < \infty$, $d \leq 1$, $d = 0$), the at-maximum indicator r, and $b_i = \beta + \tau_c z_{\mathrm{comp}(i)} + \tau_l w_{\mathrm{lin}(i)}$ with standard normal compartment and lineage deviations. A random intercept would cancel within a pool, so the random effects act on the slope. The priors, fixed before any fit, are $\alpha \sim N(0, 2.5^2)$, $\beta \sim N(0, 1.5^2)$ and $\tau \sim$ half-normal$(0, 1)$. e^{β} is the conditional odds multiplier for a candidate that attains the optimum against an otherwise identical one in the same pool, at zero compartment and lineage deviations;

Model	β median	95% interval	$P(\beta > 0)$	e^{median}	Table-1 analysis
`main`	−0.43	(−1.94, 1.11)	0.28	0.65	−0.34 (−1.84, 1.24)
`comp`	0.15	(−1.01, 1.37)	0.61	1.16	0.26 (−0.89, 1.44)
`compr`	−0.01	(−1.44, 1.37)	0.49	0.99	0.74 (−0.61, 1.96)
`pooli`	−0.40	(−1.99, 1.22)	0.30	0.67	−0.36 (−1.99, 1.27)
`compi`	0.37	(−0.97, 1.70)	0.72	1.44	0.34 (−0.98, 1.70)
`compri`	−0.23	(−1.69, 1.20)	0.37	0.79	0.98 (−0.46, 2.27)
`lineage`	0.12	(−1.04, 1.34)	0.59	1.13	0.27 (−0.90, 1.45)

Table 6: Hierarchical posterior of the robustness coefficient over the sampled histories (specification `reach`; `compr` and `compri` with the latent-state mixture likelihood under the default priors; their fixed-state fits are in Table **??**). Table-1 analysis: the same model with every origin scored as a change from table 1 with present-day decoders, median (95% interval).

β is the centre of the distribution of slopes, not the mean of the four realised compartment slopes, and not the multiplier in every pool. The first 20 of the 200 histories, which the sampler writes in random order, each get a full fit (4 chains of 1000 draws after 1000 of warm-up), and the fits are mixed with equal weights as a cut posterior that does not feed back into the history weights (Table 6).

No global interval excludes zero under any model. Under `main` β is −0.43 (−1.94, 1.11) with $P(\beta > 0) = 0.28$; under `compri` it is −0.23 (−1.69, 1.20) with the mixture likelihood and 0.97 (−0.61, 2.37) at a fixed state ($P(\beta > 0) = 0.90$). Three uncertainties are reported apart: the Monte Carlo error within fits is at most 0.010 in any quantile, the Monte Carlo error from the finite set of histories, from a bootstrap over them, at most 0.026, and the between-history share of the posterior variance lies between 0.007 and 0.043. Every sampled change (`main` −0.50 (−1.97, 1.02), `compri` 0.96 (−0.56, 2.34) at a fixed state), the ordered and unlinked layouts and the other sensitivity samplers (`main` medians −0.49 to −0.23; `compri` at a fixed state 0.75 to 1.09, merged class P 0.0011 to 0.094) give global intervals that include zero under every sampler, while the fixed-state restrictive class test ranges from rejection to non-rejection across them (Table **??**). The NCBI-taxonomy sampler satisfied the derived-reader constraint in only six draws on its unresolved Sphaeropleales and is reported but not fitted. A low between-history share says that history variation within this sampler contributes little variance; it does not say that the sampler's assumptions are right.

Contrary compartment results. The global coefficient centres a distribution over four compartments, and compartment slopes exclude zero in both directions (Table 7). In the sampled-history fits 9 compartment intervals exclude zero, 6 positive and 3 negative: under the restrictive repertoires at a fixed state the mitochondrial slopes (`compri` 2.35 (1.02, 3.86)) and the `compri` bacterial slope (1.79 (0.25, 4.00)) are positive, and under the pooled models the nuclear slope is negative (`main` −1.84 (−3.49, −0.24)). With the mixture likelihood no restrictive compartment interval excludes zero (`compri` mitochondrial 0.32 (−1.19, 2.08)). The class-conditional tests within compartments agree: at a fixed state the `compri` excess lies in mitochondria, and the homogeneity test across compartments gives merged $P = 2.0 \times 10^{-4}$ there and 0.35 under the cell null (Table **??**); no other model rejects homogeneity (`main` 0.58; Table **??**). The data favour slope heterogeneity: $P(\tau_c > 0.5)$ is 0.94 under `main` against 0.62 a priori (posterior-to-prior odds ratio in the cut mixture 10.20), and 0.95 under `compri` at a fixed state (odds ratio 12.55). These are exploratory, model-dependent associations with four compartments; they are not evidence of selection in any one compartment, but they rule out describing the results as an absence of association.

Design	Model	Decoding state	Compartment	Slope (95% interval)
catalogue changes	`compr`	fixed state	mitochondrial	1.38 (0.24, 2.89)
catalogue changes	`compri`	fixed state	mitochondrial	2.35 (1.02, 3.86)
catalogue changes	`compri`	fixed state	bacterial	1.79 (0.25, 4.00)
catalogue changes	`main`	–	nuclear	−1.84 (−3.49, −0.24)
catalogue changes	`pooli`	–	nuclear	−2.21 (−3.88, −0.56)
all changes	`compr`	fixed state	mitochondrial	1.28 (0.23, 2.61)
all changes	`compri`	fixed state	mitochondrial	2.22 (0.95, 3.63)
all changes	`compri`	fixed state	bacterial	1.69 (0.23, 3.71)
all changes	`main`	–	nuclear	−1.85 (−3.51, −0.25)

Table 7: Every compartment slope of the sampled-history fits whose 95% interval excludes zero. Catalogue changes: the primary design; all changes: every sampled change. Every compartment interval of every fit is in Tables **??** and **??**.

End-to-end calibration. Each replicate draws a true history from the sampler's posterior on the observed data (one of 100 histories, with its node states); for the restrictive generators it draws the decoding states from the default priors, independently for each reading or once per reading and shared by every transition; it draws every target from the preference model with known β and compartment slopes and with accessibility coefficients fitted to the observed data on the first truth history, as MAP estimates under the model's priors: at the no-preference fit or, for the comparisons at each tested preference, refitted with β fixed at that value (the constrained fit; Table **??** gives the coefficients and their range over the first ten truth histories); and it writes the simulated targets onto the carrier tips, rows, node constraints and reader rules. A tip takes the new target exactly when the change it inherits on the true history is retargeted, so every simulated tip state is the end state of the true history with its targets replaced (0 mismatches in 6975500 tip checks). The targets are drawn on the true history's predecessor codes and are not redrawn when an earlier change of the same codons is retargeted: 80329 of 266716 units in the 3500 audited replicates are drawn on a predecessor code that the retargeting has changed, and 2226 later changes become no-ops (Table **??**). This frozen-skeleton generator is therefore a conditional retargeting experiment on the true history's event skeleton. A second, sequential generator removes that mismatch between units: it processes each true history from the root, and each simulated unit chooses its target, or target tuple, on its simulated predecessor code, with the candidate pool, the accessibility classes, the co-occupancy and the at-maximum outcomes rebuilt on that code after every earlier simulated unit, so that the preference acts on the outcomes of the simulated transitions themselves; unresolved changes on one branch remain one joint product-pool draw, and a reader-ordered change sees the change it follows; every generated transition is checked to be admissible on its generated predecessor and every simulated tip to follow from the generated history (no failures). It uses the same random numbers as the frozen-skeleton generator, so a unit whose predecessor code is unchanged gets the same target and the two are compared in pairs. The analysis side rebuilds every pool, class and outcome from the reconstructed histories under either generator. The history sampler is then rerun, unchanged, on the simulated tips, writing 10 histories; the catalogue design and the decoding-state relevance are rebuilt; and every component exact test (exact for its fixed history) and merged P is recomputed, together with the same tests on the true history and, for a state generator, at the true state. In the first replicates of a cell the hierarchical posterior is fitted on 2 reconstructed histories (4000 draws each, the same fitting and acceptance rules). The empirical analysis mixes 20 histories, so these are calibrations of a two-history approximation to it. On the observed data the approximation was checked directly: across all 190 two-history mixtures of the 20 fitted histories, under each of the 9 fitted models, the median of β moves by at most 0.27 and an end of its 95% interval by at most 0.31 from the 20-history mixture, and no pair changes

whether the interval contains zero (0 of the pairs) or the decision $P(\beta > 0) > 0.975$ (0; Table **??**). This checks the approximation on the observed data only; it does not show that the two- and twenty-history procedures have the same coverage or power on simulated data. The exercise is a calibration conditional on the observed trees and topology alternatives, reader classes, non-carrier observations and the changes the catalogue records, under the stated generators; it does not test that the fitted history model is right. Rates are reported with Clopper–Pearson intervals (Table **??**). No sampler replicate failed (0); individual hierarchical fits that failed the diagnostics are reported below. The observed statistics of the calibrated comparisons are computed by the same procedure, the sampler rerun on the observed data writing 10 histories, so that observed and simulated statistics come from the same reduced procedure.

- **Size.** With independent decoding states and no preference (120 replicates), the exact cell null rejects at the $2\bar{P}$ rule in a proportion 0.017 (0.002–0.059) and the worst-state bound in 0.000. The fixed-state test, whose assumed state is now usually wrong, rejects in 0.100 (0.053–0.168), and the earlier prior mixture in 0.050. With states shared across transitions (100 replicates) the cell null rejects in 0.010, the fixed-state test in 0.130 and the prior mixture in 0.120 (0.064–0.200): the failure of the mixture that Proposition 10 predicts. With the state known (40 replicates) the fixed-state test rejects in 0.050, and under the pooled generator with no preference (60) the pooled analysis rejects in 0.033 (0.004–0.115).

- **Misspecified accessibility.** Under the pooled generator with no preference the fixed-state `compri` test rejects in 1.000 of replicates (0.940–1.000) and the prior mixture in 0.967, while the cell null rejects in 0.033; the uniform test rejects in 1.000. The observed statistics are not unusually small under this generator: the share of its replicates whose mean per-history P is at most the observed one (the lower-tail rank of the observed statistic) is 0.95 for the uniform test and 0.95 for the fixed-state `compri` test. Under the restrictive generator with independent states the same ranks are 0.050 and 0.033, and the cell null's is 0.64 (Table **??**). Pooled accessibility with no preference therefore generates both kinds of significant excess. This shows that neither is specific evidence of selection; it does not show that pooled accessibility generated the data, nor that it matches their magnitudes jointly.

- **Power and the calibrated detection rules.** The $2\bar{P}$ rule is conservative, and merging over reconstructed histories costs power: under the pooled generator at odds e the pooled analysis rejects in 0.200 of replicates against 0.500 with the true history known. A calibrated detection rule avoids most of that loss. Its statistic is the mean per-history P, and its critical value is the empirical 5% quantile of the same statistic over the no-preference replicates of the generator; it targets a nominal level of 0.05, and its attained size depends on that finite simulation, on ties and on the quantile rule. On fresh no-preference replicates with disjoint seeds (100 per generator) the rule rejects at 0.060 (0.022–0.126) for the cell test under the restrictive generator and 0.030 (0.006–0.085) for the pooled test under the pooled one (Table **??**). Table **??** gives the power of both tests at odds 1.5, 2, e and 3. With independent decoding states the calibrated cell test detects a preference of odds 2 in 0.450 of replicates (0.371–0.531) and of odds e in 0.725 (0.561–0.854), against 0.275 and 0.500 for the $2\bar{P}$ rule; at odds 1.5 it detects 0.250. Under the pooled generator the calibrated pooled test detects odds e in 0.275 and odds 3 in 0.400. Neither calibrated detection rule rejects on the observed data (lower-tail ranks 0.64 for the cell test under the restrictive generator and 0.69 for the pooled test under the pooled generator). Under these generators the cell test would have detected a common preference of odds e in about three replicates in four, and the pooled test in about one in four.

- **The sequential generator.** Under the sequential generator with no preference (140 replicates each) the calibrated detection rule, with the critical values of the frozen-skeleton

null cells, rejects in 9/140 replicates (0.064) for the cell test under the restrictive generator and 6/140 (0.043) for the pooled test under the pooled one, and pooled accessibility with no preference again makes the fixed-state `compri` and uniform merged tests reject in 140/140 and 140/140 replicates (interval 0.974–1.000; Table **??**, panel b). The sequential null runs reuse the seeds of the original and validation frozen null cells, so they check how the frozen critical values transport to the sequential generator rather than giving a fresh, independent validation. With accessibility refitted on the first truth history at each tested preference and then used in the sequential generator, the upper-tail ranks are 0.099 at odds 1.5 (100 replicates), 0.025 at 2, 0.025 at 2.5 and 0.010 at 3 for the restrictive cell test and 0.059 at 2 (100 replicates), 0.025 at 2.5 and 0.005 at 3 for the pooled test, against 0.109, 0.0498, 0.020, 0.015, 0.089, 0.045 and 0.010 for the paired frozen-skeleton replicates (Table **??**: intervals and seed ranges in panel a, the generator's audit in panel c: 34,035 of 112,897 units were drawn on a predecessor code that earlier simulated units had changed, every check of admissibility, code agreement and tip agreement gave zero failures, and the 205 carrier rows whose codons received different simulated targets fall outside the row check and are counted separately). At six of these seven values the sequential rank is below the paired frozen-skeleton rank, and at restrictive odds 2.5 it is 0.025 against 0.020, within simulation error; at restrictive odds 2 the rank moves from 0.0498 to 0.025, and at pooled odds 2.5 from 0.045 to 0.025. With $k = 4$ of $n = 200$ the interval for k/n reaches 0.05, so these are ranks below the line, not a precise threshold; at pooled odds 2 the rank, 0.059 ($k = 5$ of 100, interval 0.016–0.113), lies just above it.

- **Pointwise comparisons with specified preferences.** The frozen-skeleton replicates at each odds of preference also give a pointwise comparison of that preference with the observed data: under a generator with common odds ψ, the share of replicates whose mean per-history P is at least the observed one (the upper, or *exclusion*, tail), reported as $(k + 1)/(n + 1)$, measures how unusual the observed statistic would be under that generator. For each ψ the accessibility coefficients are refitted with β fixed at $\log \psi$, so that the nuisance parameters are estimated under the hypothesis being compared (Table 8; the comparisons with accessibility at its no-preference fit, and every count, interval and seed range, are in Table **??**). Under the restrictive generator with independent states the cell test gives 0.109 at odds 1.5, 0.0498 at 2, 0.020 at 2.5, 0.005 at e and 0.015 at 3; with states shared across transitions 0.026 at 2 and 0.007 at 3; and under the pooled generator the pooled test gives 0.089, 0.045, 0.025 and 0.010 at odds 2, 2.5, e and 3. On the frozen skeleton the observed statistic therefore has ranks below 0.05 at the tested common preferences of odds 2.5, e and 3 (restrictive, independent states), 2 and 3 (shared states) and e and 3 (pooled); restrictive odds 2 and pooled odds 2.5 lie near 0.05, and at restrictive odds 1.5 and pooled odds 2 the ranks are well above it. The sequential generator (previous item) gives the values reported in the abstract; the shared-state and odds-e comparisons were made on the frozen skeleton only (pooled odds 1.5 was compared only with accessibility at its no-preference fit, Table **??**). The observed statistic itself depends on the sampler's seed: over 20 seeds it ranges from 0.71 to 0.76 for the cell test and from 0.64 to 0.69 for the pooled test, against 0.74 and 0.67 over all 200 histories (Table **??**). The comparisons move with it: at the boundary the restrictive cell test at odds 2 ranges from 0.045 to 0.0498 and the pooled test at odds 2.5 from 0.040 to 0.0498 over those seeds, while the values away from the boundary keep their side of 0.05. These are pointwise results at the tested values, under generators with a common slope, fitted accessibility and the sampler's history model, and they are not monotone in ψ within simulation error; they are not a continuous bound, and they say nothing about heterogeneous compartment slopes, which were not simulated. They are ranks under fitted conditional generators; the repeated-sampling coverage of a procedure that excludes preference values after re-estimating histories and accessibility has not been established, and the size validation above concerns no-preference generators only.

		Odds of preference under the generator				
Generator	Test	1.5	2	2.5	e	3
restrictive, independent states	cell null	0.109	0.0498	0.020	0.005	0.015
	`main`	0.168	0.065	0.025	0.010	0.015
restrictive, shared states	cell null	–	0.026	–	–	0.007
pooled	`main`	–	0.089	0.045	0.025	0.010
	cell null	–	0.0495	0.030	0.020	0.005

Table 8: Pointwise comparisons of tested common preferences with the observed data: the upper-tail rank $(k+1)/(n+1)$, the share of replicates under a frozen-skeleton generator with that odds of preference whose mean per-history P is at least the observed one, with accessibility refitted on the first truth history under each preference (200 replicates per odds, 150 with shared states, 100 at odds 1.5 and at pooled odds 2). A small value gives the tested preference a small upper-tail rank under that generator; these are ranks under fitted conditional generators, without established exclusion coverage. A rank that three decimals would round to 0.050 is given to four. Counts, intervals and seed ranges: Table **??**; the sequential generator: Table **??**.

- **Heterogeneity.** With random compartment slopes around a zero centre ($\tau_c = 1$) the cell null rejects in 0.125. Its null, that every slope is zero, is false there, so this is power against a sharp null, not a failure of size.
- **Hierarchical posterior.** Under the restrictive generator with no preference the 95% interval of the mixture-likelihood fit covers β at a rate 0.967 (60 replicates, 0.885–0.996) and $P(\beta > 0) > 0.975$ occurs in 0.017 (0.000–0.089); at the fixed state the corresponding rates are 0.950 and 0.050. Under the pooled generator with no preference the pooled model covers at 1.000 (0.940–1.000), but the restrictive mixture fit covers at only 0.850 (0.734–0.929) and gives $P(\beta > 0) > 0.975$ in 0.150 (0.071–0.266): like the fixed-state test, the restrictive hierarchical model turns accessibility misspecification into a positive coefficient. At odds e the hierarchical models reach $P(\beta > 0) > 0.975$ in 0.175 (restrictive) and 0.100 (pooled) of replicates, so the absence of a global interval excluding zero says little about preferences of that size. With random compartment slopes around a zero centre the restrictive mixture fit covers at 0.900 (0.763–0.972). Each replicate's two fits are mixed only if they pass the diagnostics; some component fits did not (for example 113 of 120 accepted under the restrictive null), and a replicate with one accepted fit uses that fit alone (Table **??**).

A restricted calibration on the six merge histories, which reruns only their merge step, has more replicates and agrees where the two overlap (Table **??**; misspecification rejection 0.974 by the $2\bar{P}$ rule).

Ascertainment. The catalogue holds reassignments that were noticed. Suppose a unit whose resulting code attains its maximum is ρ times as likely to be recorded as one that does not (a ratio of recording probabilities, not an odds ratio), independently across units and of the class; recording is assumed to attach to units, although discovery usually attaches to genomes or to individual reassignment claims. Then, conditional on being recorded, the law of the statistic in each exact test is the exponential tilt of its null with $\theta = \log \rho$, by the argument of Proposition 10 applied to the recorded units. The central 95% tilt interval of each test is therefore the non-rejection set of the corresponding two-tailed inversion, at 2.5% in each tail. Separately, two values summarise the merged rules (Table **??**): the smallest ratio ρ^* at which the one-sided 5% merged reference rule stops rejecting, and the ratio $\hat{\rho}$ that matches the mean observed count under the tilt. These are tipping points of the merged reference statistics under unit-level ascertainment; the merged rules themselves are only simulation-assessed, and ascertainment was

not simulated together with reconstruction. For the uniform excess $\rho^* = 1.85$ by the $2\bar{P}$ rule (1.45 by the largest P) and $\hat{\rho} = 3.55$; for the fixed-state `compri` excess $\rho^* = 1.41$ (1.06) and $\hat{\rho} = 2.97$. Nothing in the construction of the catalogue fixes ρ; the comparison says how much outcome-dependent detection the results could absorb, not how much there was. Restricting the catalogue to units whose rows are all confirmed or supported (32–50 units per history) removes reassignments known only from genome screens. The uniform excess remains (23.9 against 11.9, merged $P = 4.1 \times 10^{-4}$), the fixed-state `compri` test gives $P = 0.0037$ and its cell null 0.30, and the hierarchical coefficients are `compri` 1.16 (−0.70, 2.79) at a fixed state and 0.25 (−1.44, 1.92) with the mixture likelihood, and `main` −0.11 (−2.01, 1.79) (Table **??**).

Equivalence. A preference was declared practically negligible, before any round-2 test, when e^{β} lies between 1/1.5 and 1.5; on these designs that margin corresponds to 3.0–4.1 extra optimum-attaining units in expectation. Equivalence is not demonstrated under any model (Table **??**, with `compr` and `compri` under the cell null and the mixture likelihood). Two one-sided tests on the class-conditional tilt give merged P of at least 0.91, the posterior probability inside the margin is 0.35–0.52, and the narrowest symmetric margins the data would support have odds 3.70–5.57 (class-conditional) and 3.04–5.60 (Bayesian). The design has little power to demonstrate equivalence at this margin: with β zero in the restricted calibration it was declared in 0.000 of 4000 class-conditional replicates (95% interval 0.000–0.001) and 0.000 of 40 hierarchical ones (0.000–0.088).

Diagnostics. Every reported hierarchical fit on the sampled histories is a full fit: of 420, all pass $\hat{R} < 1.01$ (largest 1.0099), with bulk and tail effective sample sizes of at least 400 and 0 divergences; 9 passed only on a retry with a smaller step and longer warm-up, and 0 were dropped (Table **??**). On the merge histories all 306 fits, the E3c-prior fits included, pass the same criteria (largest $\hat{R}$ 1.0097; Table **??**). Fixed-effect estimates agree with an established conditional-logit implementation to 4.3e-03 (coefficients) and 2.0e-04 (standard errors), an ensemble sampler agrees with the Hamiltonian sampler within 0.07 posterior standard deviations, and simulation-based calibration over 100 replicates gives uniform ranks for β ($P = 0.46$).

Route adjustment and alternative predictors. A prespecified extension of the hierarchical model, fitted on the same 20 sampled histories, adds co-occupancy, neighbour status and the polar-requirement cost of the amino-acid change (a measure of chemical similarity) to the reach features before the at-maximum term (Tables **??** and **??**). It removes the positive restrictive coefficient: at a fixed state β moves from 0.67 (−0.71, 1.94) to −0.27 (−1.74, 1.23) under `compr` and from 0.97 (−0.61, 2.37) to −0.26 (−1.82, 1.32) under `compri`, and the mitochondrial slopes no longer exclude zero (`compri` 0.92 (−0.47, 2.55)); under `main` it is −0.45 (−1.98, 1.14). Neighbour status carries a positive coefficient in every fitted model (`main` 1.53 (0.56, 2.53), `compri` 1.58 (0.85, 2.38)), co-occupancy a positive one under the restrictive repertoires at a fixed state (`compri` 1.96 (1.05, 2.90)), and the similarity cost includes zero in every fitted model (`main` −0.41 (−1.51, 0.73), `compri` −0.39 (−1.53, 0.81)). The robustness coefficient stays identified: stacking the within-pool differences of the complete adjusted design (the reach indicators, co-occupancy, neighbour status and cost) with the at-maximum indicator, the indicator raises their rank in every one of the 20 fitted histories under `main`, `compr` and `compri` at a fixed decoding state, and 5–9 units per history under `compri` have two candidates that agree on every added feature and differ in the optimum. This is a fixed-design check; it does not cover identification under the latent-state mixture. The positive fixed-state coefficient therefore attenuates once neighbour status, co-occupancy and the similarity cost are adjusted for together (the fit does not isolate the first two), while the similarity cost's interval includes zero; these are the five fitted models (`main`, and `compr` and `compri` at a fixed state and with the mixture likelihood). If selection favoured robust outcomes partly through those routes, the adjustment removes part

of that association too, so the attenuation is not proof that accessibility caused it. The only compartment interval that still excludes zero is the pooled nuclear slope (−1.75 (−3.46, −0.09)).

Priors, subsets and compartments. On the sampled histories, with a tighter heterogeneity prior (scale 0.25) the `compri` global interval at a fixed state excludes zero, 1.05 (0.20, 1.90), and so does the fixed-effect fit, 1.02 (0.35, 1.68); with scale 2.5 it is 0.85 (−1.22, 2.62), and with the mixture likelihood every prior gives an interval including zero (`compri` −0.32 (−1.46, 0.68) at scale 0.25, −0.37 (−1.43, 0.55) with fixed effects; Table **??**). The heterogeneous model is primary because fixed-slope inference loses coverage when slopes vary by compartment, and here the data favour variation. A second, disjoint set of 20 histories (histories 21–40) agrees with the first: all 5 of 5 fits agree within two Monte Carlo standard errors, the prespecified criterion, and every interval's inclusion of zero is unchanged (Table **??**). Leaving out one compartment at a time shows where the fixed-state restrictive coefficient comes from: without the mitochondrial units it is −0.07 (−1.45, 1.36) under `compri`, against 1.19 (−0.46, 2.68) without the plastid units, 0.66 (−1.14, 2.23) without the bacterial and 1.74 (−0.11, 3.12) without the nuclear units; no global interval excludes zero with any compartment left out (Table **??**). Every one of these 900 fits passes the diagnostics of the primary fits (largest $\hat{R}$ 1.0099, 0 dropped, 15 retried). On the merge histories, graded outcomes in place of the indicator (per pooled within-unit standard deviation) give a deficit coefficient of −0.06 (−0.98, 0.99) under `main` and 0.52 (−0.41, 1.46) under `compri`, whole-graph silent change −0.44 (−1.01, 0.22) and 0.45 (−0.08, 1.03), and polar-requirement cost 0.07 (−0.48, 0.61) and 0.22 (−0.36, 0.72): directions differ across measures and repertoires.

Scoring from table 1. Scoring the same merge-history transitions from table 1, which keeps the units and changes only the ancestral codes, gives class-conditional $P = 0.73$ under `main` and 4.6×10^{-4} under `compri`. Reconstructed ancestral codes therefore weaken the restricted-repertoire excess and leave the pooled result in place, as in the companion paper's table-1 analysis (Reich, 2026a). The remaining alternative merge histories and the catalogue issues found by the literature curation change no conclusion (Table **??**).

7 Mechanism evidence and codon capture

Order of loss and gain. Old-reader and new-reader genes were scored by parsimony across 44 assessable origins: 4 place the loss of a dedicated old-reader gene on an earlier branch than the new reading (bilaterian `AUA`→Met and `AGR`→Ser, Sphaeropleales `AGG`→Ala, *Chloroparvula* plastid `AUA`→Met), none with support the reverse (1 computed call, unsupported; below), 13 have both changes on one branch, 1 has conflicting evidence and 25 give no order evidence. These calls are consistency checks of gene presence and absence against the reconstructed codes; they are not reconstructions of functional decoding. Gene absence can reflect annotation, import or replacement, and loss of a dedicated reader is not loss of a reading: in bilaterians tRNA-Ile(GAU) survives and can read `AUA` with an unmodified G34 (Matsuyama et al., 1998, p. 3368), as Yokobori et al. (2001, p. 320) proposed for mitochondria that lost tRNA-Ile(LAU), so the old reading probably passed to another reader rather than leaving the codon unassigned. The bilaterian calls agree with the unassigned-codon readings of Sengupta et al. (2007, p. 675); the *Chloroparvula* order is stated by Turmel et al. (2019, p. 1290), and for the Sphaeropleales Žihala and Eliáš (2019, p. 3004) correlate the disappearance of `AGR` with the loss of its tRNA. The one gain-first call rests on an NCBI default code that the source contradicts (Žihala and Eliáš, 2019, Fig. 2), so it is reported separately as unsupported, and four of five CUG-Ser2 yeasts keep tRNA-Leu(CAG) beside the new tRNA-Ser(CAG) (Krassowski et al., 2018, PDF p. 5). Gene content therefore shows several dedicated reader losses, not a general functional order.

Modification timing. 28 origins have a box-resident route. Enzymes were assessable on both sides of the change for one, the bilaterian AUA→Met: ALKBH1 is present in AUA-Ile *Trichoplax* and *Nematostella*. That establishes an ortholog's presence before the change, not its mitochondrial localisation, substrate or activity. NSUN3, which writes the m^5C34 that ALKBH1 oxidises, was found only in vertebrates, *Branchiostoma* and a choanoflagellate, although f^5C34 has been found in the mitochondrial tRNA-Met of a nematode and a squid (Yokobori et al., 2001, Table 1, p. 315); failure to detect NSUN3 cannot be equated with absence of the pathway. For no origin is the complete modification pathway shown to predate the change. Giving the bilaterian change the restricted mitochondrial readings plus f^5C (merge histories) moves the compri class P only from 0.0022 to 0.0030, because the observed target, Met, already occupies the AU box and attains the optimum under every repertoire; only the classes of competing targets move.

Codon capture. A composition-based disappearance proxy, scored at the reconstructed stem of each origin, is analysed in Online Resource 1 (Section **??**, Table **??**, with the validation and weighting in panels b and c). The moved sets rank slightly above random sets of their type and kind ($P = 0.093$), but not within their own synonymous family ($P = 0.88$); the result changes at the crown and with the counting convention, and the summed disappearance rank does not discriminate the capture-supported cases, whereas low ancestral usage of the moved codons separates the small literature classes ($P = 0.0061$). Weighting codon sets by their disappearance probability moves the expected number of optimum-attaining transitions from 40.1 to 39.1 under main (observed 38.1); in this six-merge-history weighting sensitivity capture weighting changes the expectation little, and it was not repeated on the 200 primary histories.

8 Discussion

What is established. The packing optimum obeys an exact arithmetic. From a block-greedy code a single-box reassignment to a co-occupant of the box always keeps the code optimal, so the outcome is partly fixed by the decoding route, and a target-choice model identifies a robustness coefficient exactly when the realised contrasts span it. When the decoding state is uncertain the class-conditional null remains exact on the cells of the state partitions for fixed pools under the stated model (state support, class-only weights, conditional independence), whatever the law of the states. Along sampled histories the reassignments attain the optimum far more often than uniform targets would, but under the stipulated unit-level recording model a recording ratio of 1.85 for units whose resulting code attains its maximum makes the merged uniform reference statistic exceed 0.05 (and 3.55 matches the mean observed count under the tilt). Within accessibility classes the excess is not detected under the pooled, compartment and lineage repertoires. Under the restrictive repertoire with identity changes it is detected when the decoding state is fixed with every relevant reading withheld, and not when the state is uncertain and treated exactly. Conditional simulation that reruns the history sampler shows that pooled accessibility with no preference generates both the uniform and the fixed-state excess, and gives the cell null low estimated rejection rates under the tested independent- and shared-state generators. Pointwise comparisons under sequential generators refitted at each tested preference give the observed statistic upper-tail ranks below 0.05 for the tested common preferences of odds 2, 2.5 and 3 under the restrictive repertoire with uncertain decoding states and of odds 2.5 and 3 under pooled wobble rules, and the paired frozen-skeleton ranks are broadly similar; separate frozen-only shared-state comparisons also give small ranks at odds 2 and 3. At restrictive odds 1.5 and pooled odds 2 the sequential ranks are above 0.05 (0.099 and 0.059). These are ranks at tested values under fitted conditional generators, not a continuous bound, a statement about heterogeneous preferences, or exclusions with established repeated-sampling coverage. The hierarchical coefficient has no global interval excluding zero under any repertoire with the primary priors (a tighter heterogeneity prior and the fixed-effect fit do exclude zero at

a fixed state), and in the two-history calibration those intervals had little power; equivalence at odds 1.5 is not demonstrated.

What is not. The analysis does not show that decoding history caused the signature, nor that selection for the optimum was absent. The pooled repertoire is a collection of documented readings, not a reconstructed ancestral apparatus, and the benchmark against documented routes reveals differing failures but establishes no model as the ancestral repertoire. Positive estimates under restrictive repertoires and the compartment slopes that exclude zero in both directions remain open. The present evidence does not distinguish historical accessibility from a robustness preference, and the reason is located: the identifying contrasts exist, but which targets were accessible at each ancestral node depends on modifications, charging identities and route-specific readers whose ancestral states are known for almost none of the transitions, and the catalogue's detection process is not known.

Relation to the mechanism literature. The cited mechanistic literature documents non-uniform accessibility and gain–loss pathways, and argues that load minimisation is unlikely to drive modern reassignments. Knight et al. (2001b, p. 311) and Yokobori et al. (2001, pp. 318, 320) describe the box-resident routes that, by Corollary 4, attain the optimum automatically, and Yokobori et al. (2001, p. 323) conclude that tRNA features restrict how the code can change. The contribution here is narrower: for one structural endpoint, an exact account of which outcomes a route fixes, and a demonstration that inference about a preference for that endpoint along reconstructed histories changes with the assumed decoding repertoire, the decoding state and the detection process. Swire et al. (2005, p. 134), Sengupta et al. (2007, p. 684) and Sengupta and Higgs (2015, PDF p. 6) argue that load differences between codes are too small to drive modern reassignments. Our endpoint is not a load, and failing to detect a preference beyond accessibility neither confirms nor contradicts their view. Swire et al. (2005) also give a positive account of target choice: mitochondrial codes change to match the amino-acid requirements of the proteins they encode, and sense-to-sense reassignments join similar, interchangeable amino acids more often than all possible pairs do (p. 137). Chemical similarity, measured by the polar-requirement cost of the change, was added to the hierarchical model here and its coefficient includes zero in every fitted model, which does not show that similarity is irrelevant; neighbour status has positive coefficients across the five fitted specifications, co-occupancy has positive intervals excluding zero in the restrictive fixed-state fits, and with them in the model the positive restrictive coefficient for the optimum attenuates. Amino-acid demand is not modelled; to the extent that it is correlated with co-occupancy or with the optimum, it could confound the fitted associations. Sengupta and Higgs (2015, PDF p. 6) add that most sense-to-sense reassignments lie within one column of the code, another non-uniform pattern that a uniform null would misread. The capture analysis does not bear specifically on stop-codon capture: the stem same-family comparison, where disappearance has to happen, does not detect enrichment under either counting convention, and the positive stem result rests largely on two-member stop pools.

Relation to robustness comparisons. Błażej et al. (2019, p. 30) already caution that robust variants need not have been selected for robustness. The present analysis shows that, for the packing optimum, the comparison against uniform targets cannot separate the two, that it is sensitive to outcome-dependent detection, and that against accessibility-aware pools the answer depends on the repertoire and decoding state assumed. Comparisons of whole variant codes should assess sensitivity to biologically structured, accessibility-aware alternatives before reading an excess of robust changes as selection, as the companion paper argues for origins scored from table 1 (Reich, 2026a).

What would decide it. For each box-resident origin, the states that decide accessibility would have to be dated relative to the change: the enzymes that make the reading (present on the ancestral side, absent, or gained on the same branch), the modification states of outgroup tRNAs (f^5C34, m^7G34, $\Psi 35$, lysidine), the charging identity of the tRNAs that changed reading, and for stops the release-factor states. A dated state shrinks the support of Proposition 10 and so can recover contrasts for the cell null. Two further measurements would help: reassignments found by screens whose detection does not depend on the outcome, which with a model of their detection could justify or estimate ρ, and predictors of target choice from amino-acid demand. Identification, power and heterogeneity would still need to be assessed as above.

Limitations.

- Branch lengths are divergence times where TimeTree dates a node and are interpolated elsewhere; time stands in for the opportunity to change. The Mk process is symmetric and ignores disappearance and ambiguous intermediates; the gamma shape is fitted per tree rather than integrated; the histories record net changes between sampled nodes.
- Most tip states are NCBI assignments, many inherited from higher ranks; the primary tip-error model treats errors as independent across tips; a default shared by relatives is explored only in a fully coupled sensitivity sampler.
- Reader ancestry is sourced for two meanings; the replacement rate is informed by few tips and the recruitment classes are a stated prior.
- The end-to-end calibration is conditional on the observed trees and topology alternatives, reader classes, non-carrier observations and the changes the catalogue records; its replicates write 10 histories and fit the hierarchical model on 2, fewer than the empirical analysis.
- The class-conditional tests assume a common slope; under compartment heterogeneity they reject the sharp null of no preference in any compartment, which is a different question from the hierarchical model's.
- Unresolved branch groups use product pools on the common ancestral code, and the ordered layout draws orders from the codon model; a shared molecular mechanism could constrain joint outcomes further.
- Accessibility models count recognition changes; they omit expression, import, competition, amino-acid demand and fitness during intermediate states (chemical similarity is tested, through the polar-requirement cost), and the modification priors for most box-resident origins are guesses.
- In the capture analysis the summed disappearance-rank statistic failed to discriminate the literature classes (low ancestral usage did, in a small sample); the analysis never observes disappearance, and the Brownian bootstrap ignores covariance between traits.
- Ascertainment is treated as a common ratio of recording probabilities for optimum-attaining changes; detection that depends on class, lineage or compartment is not covered.
- The hierarchical fits use 20 of the 200 histories.

Reproducibility

The supplementary tables (numbered S1 onwards) and the capture analysis are in Online Resource 1. All inputs are frozen with SHA-256 sums. One command, `python regenerate.py`

in `evolution/integration`, rebuilds every number and table of this paper from the frozen inputs and the sampled histories and checks every output byte for byte against its recorded sum; `--upstream` also reruns the phylogeny, composition and tRNA workstreams from their frozen data. Every number in the text is a macro written by `make_numbers.py` or `make_numbers_paper.py` in that folder.

Statements and Declarations

Funding. No funding was received for this work.

Competing interests. The author has no competing interests to declare.

Data and code availability. A reviewer archive accompanies the submission: a summary of every headline number; every analysed primary tree with its sampled branch lengths and tips labelled by the catalogue rows they carry; the per-origin evidence table, with source, evidence class, citation, DOI, page anchor and origin posterior for every row; the test, fit, diagnostic, calibration, ascertainment and sensitivity tables; and SHA-256 certificates. The catalogue, the sampled histories, the frozen inputs, the code and every output, each with its SHA-256 sum, will be deposited in a public archive on acceptance. The companion manuscript (Reich, 2026a) is available to the editor and reviewers.

Use of large language models. The author wrote the initial material, formulated and refined the hypothesis, designed the study and checked the results. Claude (Anthropic) carried out scripting and line editing under the author's direction, and Codex (OpenAI) acted as an automated critic of drafts. The author takes full responsibility for the content.

References

Barcytė D, Žihala D, Eliáš M (2025) Expanded diversity of pedinophytes provides a window into the evolution of the genetic code in organelles. PLoS Genet 21:e1011901. `https://doi.org/10.1371/journal.pgen.1011901`

Błażej P, Wnętrzak M, Mackiewicz D, Gagat P, Mackiewicz P (2019) Many alternative and theoretical genetic codes are more robust to amino acid replacements than the standard genetic code. J Theor Biol 464:21–32. `https://doi.org/10.1016/j.jtbi.2018.12.030`

Fučíková K, Lewis PO, Lewis LA (2016) Chloroplast phylogenomic data from the green algal order Sphaeropleales (Chlorophyceae, Chlorophyta) reveal complex patterns of sequence evolution. Mol Phylogenet Evol 98:176–183. `https://doi.org/10.1016/j.ympev.2016.01.022`

Grafen A (1989) The phylogenetic regression. Philos Trans R Soc Lond B 326:119–157. `https://doi.org/10.1098/rstb.1989.0106`

Haag S, Sloan KE, Ranjan N, Warda AS, Kretschmer J, Blessing C, Hübner B et al. (2016) NSUN3 and ABH1 modify the wobble position of mt-tRNA$^{\mathrm{Met}}$ to expand codon recognition in mitochondrial translation. EMBO J 35:2104–2119. `https://doi.org/10.15252/embj.201694885`

Knight RD, Freeland SJ, Landweber LF (2001a) Rewiring the keyboard: evolvability of the genetic code. Nat Rev Genet 2:49–58. `https://doi.org/10.1038/35047500`

Knight RD, Landweber LF, Yarus M (2001b) How mitochondria redefine the code. J Mol Evol 53:299–313. https://doi.org/10.1007/s002390010220

Kollmar M, Mühlhausen S (2017) Nuclear codon reassignments in the genomics era and mechanisms behind their evolution. BioEssays 39:1600221. https://doi.org/10.1002/bies.201600221

Krassowski T, Coughlan AY, Shen XX, Zhou X, Kominek J, Opulente DA, Riley R et al. (2018) Evolutionary instability of CUG-Leu in the genetic code of budding yeasts. Nat Commun 9:1887. https://doi.org/10.1038/s41467-018-04374-7

Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G, Hedges SB (2022) TimeTree 5: an expanded resource for species divergence times. Mol Biol Evol 39:msac174. https://doi.org/10.1093/molbev/msac174

Marshall AW, Olkin I, Arnold BC (2011) Inequalities: theory of majorization and its applications, 2nd edn. Springer, New York. https://doi.org/10.1007/978-0-387-68276-1

Matsuyama S, Ueda T, Crain PF, McCloskey JA, Watanabe K (1998) A novel wobble rule found in starfish mitochondria. J Biol Chem 273:3363–3368. https://doi.org/10.1074/jbc.273.6.3363

Nakano S, Suzuki T, Kawarada L, Iwata H, Asano K, Suzuki T (2016) NSUN3 methylase initiates 5-formylcytidine biogenesis in human mitochondrial $tRNA^{Met}$. Nat Chem Biol 12:546–551. https://doi.org/10.1038/nchembio.2099

Noutahi E, Calderon V, Blanchette M, El-Mabrouk N, Lang BF (2019) Rapid genetic code evolution in green algal mitochondrial genomes. Mol Biol Evol 36:766–783. https://doi.org/10.1093/molbev/msz016

Reich J (2026a) Decoding accessibility and constrained optimality in natural codon reassignments. Manuscript.

Reich J (2026b) Transition-pair structure and near-extremal third-position transition robustness in the standard genetic code. Math Biosci, article 109813. https://doi.org/10.1016/j.mbs.2026.109813

Sengupta S, Higgs PG (2015) Pathways of genetic code evolution in ancient and modern organisms. J Mol Evol 80:229–243. https://doi.org/10.1007/s00239-015-9686-8

Sengupta S, Yang X, Higgs PG (2007) The mechanisms of codon reassignments in mitochondrial genetic codes. J Mol Evol 64:662–688. https://doi.org/10.1007/s00239-006-0284-7

Shulgina Y, Eddy SR (2021) A computational screen for alternative genetic codes in over 250,000 genomes. eLife 10:e71402. https://doi.org/10.7554/eLife.71402

Swire J, Judson OP, Burt A (2005) Mitochondrial genetic codes evolve to match amino acid requirements of proteins. J Mol Evol 60:128–139. https://doi.org/10.1007/s00239-004-0077-9

Turmel M, Lopes dos Santos A, Otis C, Sergerie R, Lemieux C (2019) Tracing the evolution of the plastome and mitogenome in the Chloropicophyceae uncovered convergent tRNA gene losses and a variant plastid genetic code. Genome Biol Evol 11:1275–1292. https://doi.org/10.1093/gbe/evz074

Yokobori S, Suzuki T, Watanabe K (2001) Genetic code variations in mitochondria: tRNA as a major determinant of genetic code plasticity. J Mol Evol 53:314–326. https://doi.org/10.1007/s002390010221

Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–314. https://doi.org/10.1007/BF00160154

Žihala D, Eliáš M (2019) Evolution and unprecedented variants of the mitochondrial genetic code in a lineage of green algae. Genome Biol Evol 11:2992–3007. https://doi.org/10.1093/gbe/evz210

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ReichReassignmentEvolutionESM.pdf