

Revisiting Xia's Index of Substitution Saturation, Part II: Iss.c is a power boundary, not a saturation threshold

Zhi-Jie Xu

zjxmolls@outlook.com

Donghua University <https://orcid.org/0009-0003-3855-0921>

Research Article

Keywords: substitution saturation, index of substitution saturation, statistical power, partition filtering, quartet, phylogenetic signal

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11438725/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Revisiting Xia's Index of Substitution Saturation, Part II: $I_{ss.c}$ is a power boundary, not a saturation threshold

Zhi-Jie Xu^{a,*}

^a*College of Biological Science & Medical Engineering, Donghua University, Shanghai 201620, China*

Abstract

We give an exact account of $I_{ss.c}$, Xia's critical value for substitution saturation, and test it in a 20,000-fragment simulation. $I_{ss.c}$ is used as a decision rule: when I_{ss} exceeds $I_{ss.c}$, the data are declared saturated and discarded. The rule cannot carry that weight.

For a four-taxon tree the critical tree is fixed by $D(T_L)\sqrt{L} = z\sqrt{v_2(T_L)}$, with D and v_2 exact sums over the 256 site patterns. Three consequences follow, without simulation. First, $I_{ss.c}$ does not measure the quantity that decides recoverability: permuting taxa within columns preserves I_{ss} exactly, to 3×10^{-15} , while destroying the signal, and along that family the recovery rate sweeps from 0.71 to 1.00 to 1/3 with I_{ss} fixed. Second, the length dependence is not a power law with a universal exponent; it is $\lambda_{I_{ss}}/(2\lambda_D - \lambda_{v_2})$, a ratio of two local rates that ranges from 0.061 to 0.585 across configurations. Third, for a subset of a concatenation it is unreliable for the decision asked of it.

Running through all three is an identity: D is exactly the Kullback–Leibler distance from the tree to its own best star approximation. The Farris zone is therefore not a separate mechanism but the region where that distance is smallest; over a controlled family D moves by a factor of 3583 while I_{ss} moves by 2.6%. In the local limit the boundary reduces to one number: $I_{ss.c}(L)$ is the index of the tree whose KL distance to its own best star approximation is $5.41/L$.

At genome scale the account is tested. Resampled 500 bp fragments from one 16-taxon tree, banded by I_{ss} at thresholds measured on that same tree, reconstruct it exactly once concatenated to 10 kbp; above 100 kbp no band differs from any other. The threshold sits where recovery has fallen six points from its peak.

Keywords: substitution saturation, index of substitution saturation, statistical power, partition filtering, quartet, phylogenetic signal

1. Introduction

Xia et al. (2003) introduced the index of substitution saturation I_{ss} and its critical value $I_{ss.c}$, tabulated by sequence length and number of taxa. Part I (Xu, 2026) showed that I_{ss} is an average over sites and cannot depend on L except through finite-sample noise, so $\text{Var}(I_{ss}) = \Theta(L^{-1})$ exactly and the critical value must move with L to hold a decision level fixed.

That leaves open what $I_{ss.c}$ is. It is a number that moves with L , on a scale whose ceiling depends on the entropy convention; it is used as a discard rule; and, as we show here, it does not measure the thing that determines whether a topology can be recovered.

*Corresponding author.

The paper is organised around three constructions, all exact and all reproducible from the accompanying scripts: a family of data sets with identical I_{ss} whose recovery rates span the whole possible range (§3.2); matched- I_{ss} genuine trees whose recovery differs by up to 125-fold (§3.3); and an exact decomposition of the length exponent into two measurable rates, with a one-factor-at-a-time account of which tree geometry moves which rate (§3.4).

Section 3.1 states the practical consequence first, because it follows from the definition of the criterion rather than from any of the machinery that follows, and it governs how the index should be used.

One result deserves flagging before the machinery, because it reorganises the subject. The Farris zone and the length dependence of $I_{ss,c}$ are usually discussed separately — one as a property of trees, the other as a property of the index, and in different literatures. They are the same quantity seen twice. The per-site gap D that sets the critical tree is, exactly, the Kullback–Leibler distance from the tree to its best star approximation; the Farris zone is the region of tree space where that distance is smallest. Section 3.5.1 establishes this and quantifies it, and it is why the factor that dominates the length exponent (§3.4) is the same one that dominates the zone.

Where this sits.. Two literatures are adjacent to this question and neither has this object. One critiques I_{ss} and its relatives on the ground of model violation — non-homogeneity, compositional heterogeneity, the treatment of invariant sites — and responds by building better statistics and calibrating them by simulation (Jermiin and Misof, 2020; Duchêne et al., 2022). Those critiques are *conditional*: their common form is that the index fails when the model is wrong. The other derives closed forms for experimental design — how much data a quartet needs, what substitution rate is optimal — without ever acting on a diagnostic statistic. Townsend (2007) derives the optimal rate for a dated four-taxon polytomy, which is to say a star, and Townsend et al. (2012) extend that to a signal-and-noise decomposition of a data set’s power to resolve a phylogeny. The gap this paper occupies is the join: the analytic objects of the second literature applied to a statistic of the first.

The claim made here is correspondingly *structural* rather than conditional. The index fails to measure what its name says even when the model is exactly right and the usage exactly as prescribed: §3.5.1 exhibits two configurations whose per-branch multiple-substitution profile is identical to the last digit and whose D differs by a factor of 36. No accumulation of conditional criticisms yields that statement, and it does not depend on any of their conditions.

Positioning that connection is worth one paragraph, because it places the paper in a debate it would otherwise appear to be ignoring. The Farris zone was named by Siddall (1998) for the four-taxon case in which the long branches are sisters, parsimony recovers the tree and likelihood does not — as against the zone of Felsenstein (1978), where the long branches are separated and parsimony itself is inconsistent. The two are not symmetric in kind: the first is a statement about consistency and holds however much data are supplied, the second is a statement about power and dissolves as data accumulate. That asymmetry is what makes the Farris zone’s boundary length-dependent, and it is the reason Leuchtenberger et al. (2020) call the zone “not unambiguously defined for ML inference” and simulate the whole branch-length plane instead of locating a boundary. Equation (3) is that boundary. The argument of Swofford et al. (2001) against Siddall (1998) — that parsimony succeeds there through its own bias rather than because likelihood is repelled — is therefore the same

claim as ours in prose; what is new here is the closed form, and with it the demonstration that the moving threshold in Xia’s tables is that boundary and not a saturation threshold at all.

The paper closes on that boundary at a scale where it can be spent. §3.6 bands twenty thousand resampled short fragments by I_{ss} and asks whether the band a filter would discard is in fact unusable, and whether a method immune to long-branch repulsion benefits from the filter; the answers are no and no, and the second answer has a mechanism that §3.6.2 computes exactly. Throughout, $I_{ss.c}$ is defined by our own resampling on our own tree, and the tabulated values are used only as a statement about which factors matter.

2. Methods

2.1. Three topologies, three likelihoods, exactly

A quartet has three candidate topologies and therefore three likelihood functions. Each is a sum of L independent per-site terms, so (ℓ_1, ℓ_2, ℓ_3) is jointly asymptotically normal and

$$\Pr[\text{correct}] = \Pr[\ell_1 > \ell_2 \wedge \ell_1 > \ell_3] \quad (1)$$

is a bivariate normal orthant probability whose means, variances and correlation are sums over the 256 site patterns. No simulation is required and no closed-form branch-length maximum-likelihood estimate is needed — none exists.

Lemma 1 (Star projection). *Both wrong topologies of a quartet have the same Kullback–Leibler projection, the star: at zero internal branch the tree is the same regardless of which split it claims, and the star family is described identically in all three parametrisations.*

Remark 2. The lemma is checked rather than assumed. Scanning 219 configurations by objective rather than by fitted branch length, the largest relative KL gain over the star anywhere is 2.5×10^{-6} ; see §4.3.

Proposition 3 (Two-way reduction). *Under Lemma 1 the three-way choice reduces exactly to a two-way one, and*

$$\Pr[\text{correct}] = \Phi\left(\frac{\sqrt{L}D}{\sigma}\right), \quad (2)$$

with D the per-site KL gap to the better alternative.

Proof. The two wrong fits coincide, so $\ell_2 = \ell_3$ identically and $\Pr[\ell_1 > \ell_2 \wedge \ell_1 > \ell_3] = \Pr[\ell_1 > \ell_2]$. That is a one-sided normal probability with mean $\sqrt{L}D$ and standard deviation σ . $\square$

One degenerate corner must be handled separately, and only a construction whose purpose is to remove all signal reaches it: when the three *fitted* distributions coincide, all three likelihoods are one number and the answer is $1/3$, not $\Phi(0) = 1/2$. See §4.3.

2.2. The critical tree

Define $I_{ss.c}(L)$ by resampling: the I_{ss} of the tree at which the recovery rate reaches 95%. This is a definition rather than a fit, needs no external data, and makes the question one of internal consistency.

Theorem 4 (Resolution condition). *Let D be the per-site Kullback–Leibler gap between the true tree and the better of the two alternatives, and v_2 its per-site variance. The tree length at which the recovery probability falls to the level z of a one-sided test, $z = \Phi^{-1}(1 - \alpha)$, satisfies*

$$D(T_L) \sqrt{L} = z \sqrt{v_2(T_L)}. \quad (3)$$

Proof. By (2) the recovery probability is $\Phi(\sqrt{L} D/\sigma)$, so the level is reached when $\sqrt{L} D/\sigma = z$. The denominator is $\sigma^2 = v_2$ by definition, which gives (3). $\square$

Corollary 5 (Exponent of the length dependence). *Writing $\lambda_\bullet$ for the local log-derivative with respect to the tree scale, the critical tree and the index at it satisfy*

$$\frac{d \log T_L^*}{d \log L} = \frac{1}{2\lambda_D - \lambda_{v_2}}, \quad I_{ss.c}(L) = I_{ss,\text{sat}} - C L^{-p}, \quad p = \frac{\lambda_{I_{ss}}}{2\lambda_D - \lambda_{v_2}}. \quad (4)$$

Proof. Differentiate (3) with respect to $\log L$. With $r = d \log T_L^*/d \log L$, the left side contributes $\sqrt{L}(\lambda_D r + \frac{1}{2})$ and the right side $D\sqrt{L} \frac{1}{2} \lambda_{v_2} r$, using $z\sqrt{v_2} = D\sqrt{L}$ from (3) itself. Equating them gives $r(\lambda_D - \frac{1}{2} \lambda_{v_2}) = -\frac{1}{2}$, that is the first relation. The index at the critical tree then moves as $d \log I_{ss.c}/d \log L = \lambda_{I_{ss}} r$, which is the third, and integrating gives the second. $\square$

The exponent is therefore a ratio of two independently measurable rates: $\lambda_{I_{ss}}$ is how fast I_{ss} saturates with tree length, and λ_D is how fast the topological signal decays — that is, the multiple-hit effect itself.

2.3. The simulation experiment

The exact results need no simulation; the genome-scale test does, and its design is fixed before any measurement is taken.

The tree is a single 16-taxon Yule–Harding tree drawn once by AliSim and kept for the whole study. Fragments are 500 bp simulated under that tree with every branch length multiplied by a common scale s , for twenty values of s spanning 0.04 to 6, so that tree length is the only parameter that varies and the topology is held fixed. Twenty thousand fragments are drawn this way, forming a pool of 10 Mbp. Because sites are i.i.d., a long alignment chopped into 500 bp pieces is exactly a set of independent 500 bp fragments, so one AliSim call per scale produces a whole block of the pool.

I_{ss} is computed per fragment, convention A, from the simulated bases. The two thresholds that partition the pool are then measured on this same tree by the resampling procedure of §2.2: a fragment is called clean by a length- L calibration when its I_{ss} lies below $I_{ss.c}(L)$. The arms are the bands between the thresholds, plus the integrated pool; each is truncated to the same number of sites, so that the arms differ only in composition and not in length.

Three methods are run on every arm: maximum likelihood under JC69 through IQ-TREE, and both CASTER variants. Recovery is scored as the Robinson–Foulds distance

to the true tree, whose maximum on 16 taxa is 26. Uncertainty is estimated by a delete-a-block jackknife over the fragment pool rather than by independent replicates, because the question is how the result depends on which fragments were drawn; the pool is split into ten consecutive blocks and each is deleted in turn.

2.4. Software and reproducibility

Simulation uses AliSim and recovery uses IQ-TREE; `caster-site` and `caster-pair` are used unmodified. The quartet results — the resolution condition, the exponent decomposition, the orthant probabilities and the closed form for CASTER’s score — are computed by enumeration and orthant quadrature and need no external binary at all.

Every number in this paper is produced by a script in the archive, one per result set, and every table is read from the CSV that script writes. Code and result tables are at https://github.com/Gipsy-The-Shellier/Iss_simulations and deposited at <https://doi.org/10.5281/zenodo.23470551>; the repository holds both papers of this series, and the tag `rs-preprint` is the version this manuscript was built from.

3. Results

3.1. The criterion is not valid for a subset of a concatenation

Substitution saturation is a property of the substitution process. A subsample of an alignment underwent the same process as the whole: it is a window onto the same history, not a different history. Any quantity that measures saturation — the frequency of multiple hits, the fraction of sites made homoplastic, the expected site entropy — has the same expectation on the subsample as on the whole.

An alignment of 10,000 bp may have $I_{ss} < I_{ss.c}$ while a 500 bp subalignment of it has $I_{ss} > I_{ss.c}$. This is not evidence that the subsample is more saturated. It is two effects, neither of them saturation.

1. I_{ss} is a per-site average, so $\mathbb{E}[I_{ss}]$ is the same for any subset of sites; the difference between the two observed values is estimator noise, with standard error $\Theta(L^{-1/2})$, four and a half times larger at $L = 500$ than at $L = 10,000$.
2. $I_{ss.c}(L)$ *increases* with L (§2.2). The threshold at $L = 500$ is lower than at $L = 10,000$, so the same data cross it more easily. The data did not change; the threshold did.

Therefore:

- (i) $I_{ss.c}$ has a defined decision-theoretic meaning only for the whole data set it was computed for. Applied to a partition, a gene, a single-copy locus, or any subalignment, the comparison $I_{ss} > I_{ss.c}$ no longer tests the hypothesis it names, and the resulting decision — in particular the decision to *discard* a partition as saturated — is not supported by the statistic.
- (ii) The power of the criterion is a property of the full set. A subset has less power by construction, and no reading of $I_{ss} > I_{ss.c}$ on a subset recovers it.
- (iii) Consequently the operative question for a partition is not its saturation level but whether it carries a different history — the evolutionary-heterogeneity question. Saturation is a shared property of the whole alignment’s substitution process; heterogeneity is not, and it is the one that partitioning actually addresses.

(iv) Where a method can down-weight rather than exclude, it dominates discarding. Quartet-based and supertree methods admit exactly this: weighted ASTRAL (Zhang et al., 2018) assigns each gene tree a quantitative weight, so an unreliable or conflicting partition contributes less without being thrown away, and the information it does carry is retained. A binary discard/keep rule based on $I_{ss,c}$ cannot express that, and the replacement of a graded weight by a threshold is a loss of information that no choice of threshold recovers. The empirical literature on the practice this replaces — Duchêne et al. (2022) is the most direct recent example — measures how much is lost by excluding loci, and the question is quantitative rather than binary in exactly that way.

This is not a hypothetical. Slicing an alignment into blocks and judging each block against the threshold tabulated for the block’s length gives a quantitative version (Fig. 5). The blocks are slices of the same alignment, so their expected I_{ss} is the whole’s — that is, exactly the whole’s threshold. The threshold they would be judged against, however, is the one tabulated for the shorter length, and it is lower. Across four (block, whole) pairs, spanning whole lengths from 4000 to 40,000, the block’s mean sits 2.3 to 3.0 block standard deviations above its own threshold, and between 19.8 and 20.0 of every 20 blocks are declared saturated while the whole passes.

Below some alignment length the criterion is not merely wrong but undefined. For the Farris tree of Fig. 1 no tree length reaches a 95% recovery rate at any $L \leq 10^4$ — at $L = 3000$ the best achievable rate over all tree scales is 0.755 — so there is no threshold to compare against. A tabulated $I_{ss,c}$ supplies a number at every length regardless, so in practice a verdict is always available; it is simply not a verdict the data support.

A corollary, stated carefully, because the obvious version of it is wrong. If the failure were caused by saturation, discarding sites would look rational, since more data could not help. If it is caused by power, the tempting conclusion is that discarding is backwards — deleting data to cure a shortage of data. That conclusion does not follow, and an earlier draft of this section drew it. What matters is not how many sites are removed but whether the removed ones carried signal. Filtering on site entropy — which is what the index does, one site at a time — keeps the informative sites and drops the rest, and in this framework it *helps*: measured on three configurations at their critical trees, removing the sites above entropy 0.90 raises standardised signal by 22%, 40% and 337%. The practice of discarding saturated sites is therefore not refuted by the power reading.

What the power reading does say is narrower and survives: the benefit is a property of *which* sites the filter selects and not of how many, so it is not transportable. A filter calibrated on one configuration selects a different set of sites on another, and §3.5.1 shows the index varies enough between configurations to make that the dominant consideration — at a fixed alignment length, the same index reads 0.33 on one tree and 0.99 on another. The remedy is not "keep the data" or "discard the data" but "know which sites your filter is removing", and that is a question about the configuration rather than about the index.

Finally, the correlation that makes the rule look reasonable in practice is a confounding. Short partitions often are fast-evolving genes — third codon positions, rRNA loop regions — and it is the rate, not the length, that saturates them. That is precisely why the question of §3.1(iii) is the right one to substitute.

3.2. I_{ss} does not determine recoverability

Relation to Part I. Part I (Xu, 2026) already reports that along a four-taxon transfer of long branches at fixed patristic total, the index is constant while recovery spans its full range, and that a fifteen-fold change in the tree length at which recovery fails moves the index by 1.2%. Those are the same phenomena studied here, and the present section is a strengthening rather than a discovery: Part I's invariance holds along a path that preserves the patristic total, whereas the construction below holds I_{ss} fixed *exactly*, to machine precision, at every point of a family that is not a path of trees at all. What is new in Part 2 is the mechanism (3), the exponent decomposition, the identity (8), and the consequences of §3.1; what is not new is the observation that the index fails to track recoverability.

Proposition 6 (Invariance under within-column permutation). *Permute the taxa within each column, independently per column, and let $P_f = (1 - f)P + fQ$ where Q is the average of P over all 24 relabellings. Then $I_{ss}(P_f) = I_{ss}(P)$ exactly, for every $f \in [0, 1]$.*

Proof. Each column keeps its multiset of bases, so the site entropy h is unchanged; each column contributes the same counts to the pooled composition, so the reference entropy is unchanged. That gives the case $f = 0$.

The extension to mixtures is *not* a linearity argument: I_{ss} is not linear in the site distribution, its denominator being a nonlinear function of the pooled composition. What holds is that numerator and denominator are separately invariant. Both $\bar{H}$ and $\hat{p}$ are linear, and Q has the same value of each as P does, because every relabelling preserves each column's multiset and hence the pooled composition. Hence

$$\bar{H}(P_f) = (1 - f)\bar{H}(P) + f\bar{H}(Q) = \bar{H}(P), \quad \hat{p}(P_f) = (1 - f)\hat{p}(P) + f\hat{p}(Q) = \hat{p}(P), \quad (5)$$

and the reference entropy, which depends on $\hat{p}$ alone, is unchanged. $\square$

At $f = 1$ the distribution is invariant under every relabelling, so the three topologies are exchangeable and the rate is exactly $1/3$.

Figure 1 shows the consequence on a Farris-zone quartet ($a = b = 0.05$, $c = d = 1.0$, $e = 0.02$, $T_L = 2.12$) at $L = 30,000$. The decision statistic sweeps from $z = 0.56$ to $z = 7.81$ and back to $z = 0$ while I_{ss} does not move in the fifteenth decimal. The rate is non-monotone in f because the shuffled component is unfittable by any topology and degrades the *wrong* topologies faster than the right one.

3.3. The same statement inside the tree family

The construction above uses data that are not trees. The same conclusion can be drawn from two genuine trees under the same model.

I_{ss} is monotone in tree scale within a shape, so one can solve for the scale of a balanced tree matching a Farris tree's I_{ss} exactly. At matched I_{ss} the balanced tree's standardised signal is 19 to 125 times the Farris tree's (Fig. 2). The geometry is visible directly: at $I_{ss} = 0.1727$ the Farris tree has internal branch 0.003 and the balanced tree 0.096 — a factor of 32 — though their I_{ss} agrees to five decimals.

Table 1: The exponent as a ratio of two rates, one factor at a time against the reference $a = b = 1.0$, $c = d = 0.05$, $e = 0.02$.

configuration	λ_D	λ_{v_2}	$\lambda_{I_{ss}}$	p_{pred}	p_{fit}	rel. err.
reference	3.324	3.354	0.934	0.2836	0.2897	-0.021
$e \times 5$	4.106	4.111	0.555	0.1353	0.1374	-0.016
$e \times 25$	5.094	5.104	0.312	0.0613	0.0610	+0.005
$c = d \times 4$	3.280	3.313	1.249	0.3848	0.3904	-0.015
$a = b \div 2$	1.802	1.807	0.439	0.2442	0.2489	-0.019
$a = b \div 5$	0.827	0.828	0.226	0.2733	0.2758	-0.009
all equal	1.596	1.600	0.811	0.5089	0.5085	+0.001

3.4. What sets $I_{ss,c}$, and by how much

3.4.1. Verification

Table 1 varies one factor at a time against a common reference. The prediction holds to a mean absolute relative error of 1.2% and a maximum of 2.1% over seven configurations. Pointwise, comparing the local exponent against the locally measured ratio at each L , the mean is 1.0% and one configuration reaches 0.07%; the residual is systematically negative and shrinks as the critical scale grows, consistent with a higher-order term rather than a missing mechanism.

3.4.2. A regularity worth separating

λ_{v_2} tracks λ_D with slope 1.0031 and correlation 0.999966 across all seven configurations, so $2\lambda_D - \lambda_{v_2} \approx \lambda_D$ and

$$p \approx \frac{\lambda_{I_{ss}}}{\lambda_D}. \quad (6)$$

The reason is not a coincidence of this family. In the local regime the likelihood-ratio relation gives $\mathbb{E}[\Lambda] = \frac{1}{2} \text{Var}[\Lambda]$ for close alternatives, that is $v_2 \rightarrow 2D$, so the two share a rate exactly; and the local regime is the one the resolution boundary occupies, since the critical tree is by construction the most saturated tree still resolvable. The identification is therefore exact where it is used and degrades where it is not: v_2/D is 2.0035 on the Farris trees and 3.0054 on the balanced one, rising monotonically with D (§3.5.1). Substituting it into (3) collapses the boundary to a single number, (10).

3.4.3. Which factor matters

With $p \approx \lambda_{I_{ss}}/\lambda_D$ the factor table reads directly (Fig. 3B). The internal branch dominates: multiplying it by 25 takes λ_D from 3.32 to 5.09 and $\lambda_{I_{ss}}$ from 0.93 to 0.31, so p falls from 0.290 to 0.061 — a factor of 4.7. The short pendant pair is second: multiplying it by 4 leaves λ_D flat while raising $\lambda_{I_{ss}}$ from 0.93 to 1.25, taking p to 0.390. The long pendant pair barely matters, because dividing it by 5 moves both rates together and leaves the ratio near its reference value.

3.4.4. Where the criterion ceases to exist

Of the eight configurations scanned, seven admit a measurement and one does not. The near-star tree ($e = 0.002$) has no crossing of 95% recovery at any tree length in the range

tested, so $I_{ss,c}$ is not defined for it. The all-equal tree ($a = b = c = d = 0.3$, $e = 0.6$) is the nearest thing to a second such case and does admit one, fitting to $r^2 = 1.0000$ with a relative error of 0.1%. The distinction is worth stating because the two are easily confused: a configuration whose threshold is hard to locate looks exactly like one that has none.

3.4.5. Validation against resampling

Over 16 points spanning four Farris shapes, $I_{ss,c}$ from exact multinomial resampling agrees with the analytic value: $\max |\text{residual}| = 0.00324$, $\max |z| = 1.29$, standardised residuals with mean +0.172 and bootstrap 95% confidence interval $[-0.087, +0.438]$ containing zero, median +0.057, sign test 8 of 16 positive. No systematic bias is detected (Fig. 6).

The standardised residuals have standard deviation $0.561 < 1$, so the error bars are conservative by about $1.8\times$, because a shared random seed correlates the Monte Carlo noise across scales and it partly cancels in the bisection. With 16 points the sign test has no power — 47 are needed for an 80% chance of detecting a 70/30 split — so this is a failure to detect a bias, not evidence of its absence.

3.5. Why the exponent is not a constant

Fitting $I_{ss,\text{sat}} - I_{ss,c}(L) = CL^{-p}$ over a decade gives a good straight line, but it is not exact and the exponent drifts. For one Farris tree the local exponents across $L = 3 \times 10^4$ to 10^6 are

$$0.352, 0.307, 0.264, 0.232, 0.208, 0.180, \quad (7)$$

monotonically decreasing. A single power law fits with $r^2 = 0.98688$; adding one $\log^2 L$ term lifts that to 0.999955. The exponent extracted from a window is a local average of a drifting quantity, which is exactly what $p \approx \lambda_{I_{ss}}/\lambda_D$ predicts, since both rates are local.

This is why Part I's $L^{-1/2}$ — a correct statement about the *sampling* threshold at a fixed tree — does not govern $I_{ss,c}$: that is a different object. Nor is any other single exponent available, since the shapes tested here, all of them symmetric in the sense $a = b$, $c = d$, span p from 0.061 to 0.585.

3.5.1. The Farris zone and $I_{ss,c}$ are one quantity seen twice

What the Farris zone is.. In the Felsenstein zone the long branches are separated by a short internal edge and parsimony is positively misleading regardless of how much data are supplied (Felsenstein, 1978; Hendy and Penny, 1989). In the Farris zone the long branches are themselves sisters, parsimony recovers the correct tree, and likelihood is the method that struggles; Siddall (1998) named both the phenomenon (long-branch repulsion) and the region. The two zones are not symmetric in kind. The first is a statement about consistency, and does not depend on the length of the alignment. The second is a statement about power, and does: Leuchtenberger et al. (2020) write that the Farris zone “is not unambiguously defined for ML inference because its boundary depends on the length of the sequence alignment”, and work around the difficulty by simulating the whole branch-length plane rather than locating a boundary. What follows supplies the boundary they declined to locate. Swofford et al. (2001) had already argued that parsimony's success there is a consequence of its own bias rather than of likelihood's failure, which is the same claim in words; equation (3) below is the quantitative form.

The identity..

Theorem 7 (The gap is a distance to the star). *Let P be the true site-pattern distribution of a quartet, Q_i^* its Kullback–Leibler projection onto topology i , and star^* the star onto which both wrong topologies project. The per-site gap D of (3) is then nothing other than the distance from the tree to its own best star approximation:*

$$D = \text{KL}(P_{\text{true}} \parallel \text{star}^*). \quad (8)$$

Proof. The expected log-likelihood advantage of the true topology over a wrong one, per site, is $\text{KL}(P \parallel Q_2^*) - \text{KL}(P \parallel Q_1^*)$, and D is that quantity for the better of the two alternatives. The true topology’s projection is P itself, so $\text{KL}(P \parallel Q_1^*) = 0$. Both wrong topologies project onto the same star (§2.1), so their divergences coincide and the remaining term is the stated distance. $\square$

The identity is checked numerically as well as proved: over five trees spanning both zones the ratio of the two sides differs from 1 by at most 1.2×10^{-10} .

If that is what D is, then two things that appeared to be separate phenomena are the same quantity appearing twice. The difficulty of resolving a tree is D in the numerator of the resolution condition. The Farris zone is the region of tree space where D is small, because what makes the long-branches-sisters arrangement hard is that it drives the tree *towards being a star*.

The controlled family makes this quantitative (Fig. 4). Holding the pendant sum at 4, the internal branch at 0.2 and hence the total tree length at 4.2 throughout, and moving only the arrangement:

r	D Farris	D Felsenstein	ratio	I_{ss} , Farris / Felsenstein
0.0	5.52×10^{-4}	5.52×10^{-4}	1.00	0.9320 / 0.9320
0.4	1.42×10^{-4}	5.00×10^{-4}	3.53	0.9149 / 0.9191
0.8	1.94×10^{-5}	6.95×10^{-4}	35.8	0.8416 / 0.8634

D falls by a factor of 28 along the Farris family and stays flat along the Felsenstein one, and it does so with every other tree parameter fixed. The Farris zone is thus a *region of tree space* — the region nearest the star — rather than a separate mechanism, and the quantity that defines it is the same one that sets T_L^* and hence $I_{ss,c}$.

The last column is the sharpest single statement of §3.2 available. Over this range D moves by a factor of 3583 while I_{ss} moves by 2.6%. Whatever the index is measuring, it is not the quantity that decides whether the branch can be resolved.

Why I_{ss} moves so little, and why Part I predicted it.. The 2.6% is not a numerical accident of this family, and it is not observed so much as derived. Part I (Xu, 2026) shows that under uniform base composition — which is the case here, JC69 — the index collapses exactly onto the mean pairwise p -distance. For a quartet the pairwise path lengths sum to a quantity that depends only on the branch lengths and not on the topology,

$$\sum_{i < j} d_{ij} = 3(a + b + c + d) + 4e, \quad (9)$$

so the two arrangements of the family have *identical* distance sums at every r : $3(4)+4(0.2) = 12.8$ for both. They differ only in how that total is partitioned, and the mean p -distance responds to the partition only through the concavity of $t \mapsto \frac{3}{4}(1 - e^{-4t/3})$. Measured across the family, the ratio of mean p -distances tracks the ratio of I_{ss} values to within half a percent:

r	mean p , Farris / Felsenstein	ratio	I_{ss} ratio
0.0	0.706019 / 0.706019	1.00000	1.00000
0.4	0.695164 / 0.697705	1.00366	1.00453
0.8	0.649031 / 0.662370	1.02055	1.02583

I_{ss} is therefore blind to the arrangements here for a structural reason: to leading order it is a function of the mean pairwise divergence, the pairwise totals are equal by (9), and the residual is a second-order term. That is also why Part I's transfer path found the index constant — it held the patristic total fixed.

What the boundary reduces to.. In the same regime $v_2 \rightarrow 2D$. This is the local form of the likelihood-ratio relation rather than a phylogenetic fact: for alternatives that are close, $\mathbb{E}[\Lambda] = \frac{1}{2} \text{Var}[\Lambda]$, so $v_2/D \rightarrow 2$ as $D \rightarrow 0$. Measured on the trees above, v_2/D is 2.0035 in the Farris zone and rises monotonically with D , reaching 3.0054 on the balanced tree where $D = 0.229$; it is a local statement, and the local regime is where the resolution boundary lives. Substituting into (3),

$$D(T_L^*) \rightarrow \frac{2z^2}{L}, \quad \text{that is,} \quad D(T_L^*) \rightarrow \frac{5.41}{L} \quad \text{at 95\%}. \quad (10)$$

L	$D(T_L^*)$ solved	$2z^2/L$	rel. err.	v_2/D
20,000	2.870×10^{-4}	2.706×10^{-4}	+0.061	2.122
30,000	1.850×10^{-4}	1.804×10^{-4}	+0.026	2.051
100,000	5.437×10^{-5}	5.412×10^{-5}	+0.005	2.009
200,000	2.712×10^{-5}	2.706×10^{-5}	+0.002	2.005

The error and $v_2/D - 2$ fall together, because as L grows the critical tree is pushed further into saturation, $D(T_L^*) \rightarrow 0$, and the local relation becomes exact. The statement in words is short enough to check by hand: $I_{ss,c}(L)$ is the index of the tree whose KL distance to its own best star approximation is $5.41/L$.

And the boundary is two-valued.. Equation (10) locates a level set of D , and D is *not* monotone in tree length: it vanishes at both ends of the range (for a tree shrunk to a point, and for a tree grown past saturation) and has an interior maximum. I_{ss} by contrast is monotone — it is the ratio of two entropies and cannot be anything else. So the boundary is a level set of a non-monotone function of the tree, located by inverting a monotone one, and a level set below the maximum has *two* solutions.

Both are real trees, and they carry wildly different values of the index. On the balanced family at $L = 20,000$, where the required D is 2.71×10^{-4} :

branch of the level set	T_L	I_{ss}
ascending	0.0001	0.00005
descending	9.618	0.99417

The criterion in its tabulated form names one number for a given L ; the condition that the reading of §3.5.1 reduces to admits two, and they differ by four orders of magnitude in I_{ss} . This is the mechanism behind the observation of Leuchtenberger et al. (2020) that the zone is not unambiguously defined for likelihood: not vagueness, but a two-valued level set. Choosing the descending branch — the saturated one, where a tree has to be pushed until it can no longer be resolved — is the reading adopted throughout this paper, and it is a choice about which side of a maximum one is on, not a consequence of the definition.

Precedents.. The idea that quartet resolvability is a distance from the star is not new and should not be presented as such. Strimmer and von Haeseler (1997) plot the three quartet likelihoods in a simplex whose fourth vertex is the star tree, and read the central region as star-like evolution, that is, as no resolution — the geometric picture, without a distance and without a length dependence. Su and Townsend (2015) go further and give an analytic signal-and-noise decomposition for quartets under JC that treats both the Felsenstein and the Farris zone and locates branch-length conditions where inference becomes inconsistent; their analysis is for parsimony, where ours is for likelihood, and they identify thresholds rather than a boundary that moves with L . What is added here is (3) and its consequence (10), which turn the picture into a computable boundary at every L .

This is not geometry-mediated saturation.. A natural objection is that fixing the pendant sum and the internal branch fixes only the total amount of multiple substitution, not its distribution across branches, and that the Farris arrangement concentrates that length on two sister branches and is therefore merely a locally more saturated tree. The objection has a testable consequence and it fails it.

At every r the two arrangements have the *same multiset* of branch lengths: $(1 + r, 1 + r, 1 - r, 1 - r, 0.2)$ and $(1 + r, 1 - r, 1 + r, 1 - r, 0.2)$ differ only in which leaf each length is attached to. Substitutions on distinct branches are conditionally independent given the tree, so any quantity that is a function of the per-branch counts alone is identical between them — not approximately, identically. The per-branch multiple-hit probabilities are the same sorted vector, and so is the probability that a site carries a multiple hit on some branch:

r	$P(\text{multiple hit somewhere})$	ratio of D	I_{ss} ratio
0.0	0.712085 (identical)	1.00	1.000
0.4	0.734657 (identical)	3.53	1.005
0.8	0.796847 (identical)	35.8	1.026

At $r = 0.8$ both arrangements have 79.68% of sites carrying a multiple hit somewhere, to the last digit, while D differs by a factor of 36. The local saturation is not heavier in one than the other; it is the same number.

What can differ between the arrangements is a functional of the *joint* pattern distribution, and I_{ss} is one: it moves by 2.6% across this range. A defence of the form “this is saturation

under another name” must therefore nominate a saturation statistic with the dynamic range to carry the effect. The only candidate that moves at all moves by 2.6%, against D ’s 3583%.

3.5.2. Against Xia et al.’s empirical ordering

Xia et al. report that the asymmetrical ultrametric tree’s $I_{ss,c}$ converges more slowly than the symmetrical one’s. The rate falls with asymmetry in *both* arrangements, from $p = 0.637$ at $r = 0$ to 0.158 (Farris) and 0.223 (Felsenstein) at $r = 0.8$. So the reported ordering is reproduced and explained by asymmetry, which contributes a factor of about 3; the Farris arrangement adds a further factor of about 1.4 on top, through the collapse of D described above.

What the controlled family also shows is that rate and difficulty are *decoupled*. Difficulty, read as the critical tree length at a fixed L , falls from 5.386 to 2.946 in the Farris family — the tree becomes harder — but stays flat, 5.386 to 5.506, in the Felsenstein family, while the rate there still falls by a factor of 2.9. Slowness of convergence is therefore not a consequence of hardness: both are downstream of D , and they move together only where D is the thing that moves.

3.5.3. Relation to Part I, and one correction to it

Part I (Xu, 2026) is the companion to this paper, and the two differ in more than level of detail. Four things belong on the record.

Part I’s $L^{-1/2}$ does not fix the functional form.. Part I establishes that $\text{Var}(I_{ss}) = \Theta(L^{-1})$ and concludes that “the critical value must move with L at rate $L^{-1/2}$ to hold a decision threshold fixed”, describing this as “the mechanism behind Figure 3” which “fixes the functional form of the tabulated surface”. The variance statement is correct; the inference drawn from it is not, and this paper supersedes it. $L^{-1/2}$ is the rate at which the *sampling* threshold of a fixed tree moves. $I_{ss,c}$ is not that object: the tree whose recovery probability is 95% is a different tree at each L , so the threshold is set by the motion of the critical tree as well as by the noise. §3.5 measures the resulting exponent at 0.061 to 0.585 across configurations, drifting within a single configuration, and identifies it as a ratio of two local rates rather than a constant. There is no universal exponent here, and in particular not one half.

The zone and its name.. The four-taxon case in which the long branches are sisters is named by Siddall (1998), whose title names it, and §3.5.1 and §3.6.4 follow that attribution; Hendy and Penny (1989) give the general framework for comparing trees that the formalism here uses.

Notation.. Part I writes Δ for a composition-leakage term, $\sum_{\mu \neq 0} (w_\mu - \tau_\mu/K) \xi_\mu$. The quantity named here is a different object — the distance from the tree to its own best star approximation, (8) — and it is written D throughout so that the two papers cannot be confused on a quick reading.

Part I measured something this paper should have used.. Section 2.13 of Part I reports that in the transition band $\bar{p} \in [0.63, 0.70]$ the mean pairwise distance is nearly useless, with AUC 0.555 and 0.604 for MPBoot and ML, while I_{ss} on its own reaches AUC 0.750 and 0.800, and that adding I_{ss} to $\bar{p}$ raises AUC by +0.17 and +0.18. The index therefore carries real

band	sites			RF distance at 10 kbp		
	count	Mbp	mean I_{ss}	ML	CASTER-site	CASTER-pair
CLEAN	10,949	5.47	0.166	0	0	0
MID	2,050	1.03	0.555	0	0	0
HIGH	7,001	3.50	1.042	0	2	2
integrated	20,000	10.00	0.512	0	0	0

Table 2: The band design. Fragments are 500 bp drawn from one 16-taxon Yule–Harding tree at twenty uniform rescalings of its branch lengths, so only the tree length varies and the topology is held fixed. The two cuts are measured, by the resampling procedure of §3.4.5, on the tree the fragments came from. **MID** is the band the two readings disagree about: a 500 bp calibration rejects it, a 10 kbp calibration accepts it. Every arm is truncated to the same 1.025 Mbp, so the columns differ only in which fragments they contain.

information in exactly the band where the decision is made. That is not in tension with the claim of §3.2 that I_{ss} does not measure the quantity that decides recoverability, but it is in tension with a careless reading of it, and the reconciliation is Part I’s own: what the index tracks in that band is tree shape and higher-order composition, it does so only within a few points of the ceiling, and what it carries is information about the configuration rather than about whether a particular alignment can be resolved. The finding of §3.6 — that the index identifies a real gradient but places its boundary where the gradient has barely begun — is the empirical form of that reconciliation.

3.6. The criterion at genome scale

Everything to this point is quartet algebra. This section is its empirical counterpart, and it exists to settle a question the algebra cannot reach.

Duchêne et al. (2022) report that excluding loci with high I_{ss} improves phylogenomic inference. On the reading of that practice developed here, what the filter removes is not saturation but sites that carry no topological signal, and the two readings make different predictions about a method that is not susceptible to long-branch repulsion. If the benefit is a saturation effect, it should accrue to any method. If the benefit is the repair of long-branch attraction, then a method without that defect has nothing to gain, and the filter can only remove data.

The two cuts. $I_{ss,c}$ is measured here the way §3.4.5 measures it, on the same tree, at 500 bp and at 10 kbp: 0.4377 and 0.6658. The order is the one Xia et al. (2003) report and the one the theory requires — more data tolerates more saturation — but the value that matters for what follows is the distance between them, because a fragment whose I_{ss} falls between the two is data that one calibration calls saturated and the other calls clean.

The definition of the cut has to be got right, and an earlier reading of it in this work was wrong. $I_{ss,c}$ is not where recovery climbs through the level. It is, in Xia et al. (2003)’s own words, where the sequences “begin to fail to recover the true tree” — the *falling* crossing, the largest tree length still recoverable. The distinction is not pedantic: recovery against tree length rises and then falls, since at large tree lengths all three topologies approach the same saturated distribution and D goes to zero, so (3) has two roots and only the outer one is the criterion. Reading the inner one inverts the dependence on alignment length.

3.6.1. The threshold is ill-conditioned

The cut also has to be taken with care for a reason that is a property of the index rather than of our implementation. On this tree, recovery at 10 kbp as a function of I_{ss} falls like this:

$$\begin{array}{c|ccccccc} I_{ss} & 0.72 & 0.84 & 0.96 & 1.07 & 1.16 & 1.24 & 1.30 \\ \hline \text{recovery} & 0.90 & 0.92 & 0.86 & 0.74 & 0.60 & 0.20 & 0.16 \end{array} \quad (11)$$

The measured $I_{ss,c}(10 \text{ kbp}) = 0.6658$ sits where recovery has fallen from its peak of 0.96 to about 0.90: a six-point change. The cliff is at $I_{ss} \approx 1.2$, nearly twice the reported threshold, and recovery is still 0.74 at an I_{ss} half again above the cut. A threshold placed on the shoulder of a curve is not a critical value in any operational sense; it is the location of a crossing on the part of the curve that is nearly flat, and it inherits three instabilities that this study ran into directly.

Moving the level moves the threshold. Reading 90% instead of 95% would place $I_{ss,c}(10 \text{ kbp})$ near 1.0, a change of fifty per cent in the number a user compares against.

Noise moves it too. With fifty replicates per point the estimate was not even monotone in L : $I_{ss,c}$ came out as 0.44, 0.51, 0.45, 0.28, 0.67 across 500, 1000, 2000, 5000 and 10 000 bp. Only the two endpoints are used above, and they are the two that the design needs, but the intermediate values show how little separates the signal from the noise.

And the definition itself moves. Xia et al. (2003) define TL_c as the tree length where $P_{\text{true}} \geq 0.95$ for the symmetrical topology, but for the asymmetrical one, where “there is often no tree length at which the true tree is recovered 100%”, they take the tree length where P_{true} has fallen to 95% of its own maximum. On this tree the strict level is unattainable at 500, 1000 and 2000 bp — recovery peaks at 0.72, 0.84 and 0.90 — and attainable at 5000 and 10 000 bp, so a single curve through the five lengths silently changes definition halfway along it. The two rules are not the same measure, because “95% of the maximum” means something different when the maximum is 0.72 than when it is 0.98.

None of this is an argument that the index is miscalibrated. It is an argument that its calibration is a crossing on a flat part of a curve, and that flatness is a property of the quantity rather than of the estimate.

3.6.2. The same machinery applied to another index

Before turning to the experiment, the quartet algebra can be pointed at a second published threshold. The comparison is unusually clean, because the authors of that threshold carried out the same computation themselves, and the two agree exactly.

CASTER assigns a fixed vote to each site according to the four-taxon pattern it induces: +1 to the topology a two-base site resolves, and a weight w_3 to the topology that groups the pair sharing a base at a three-base site, with one- and four-base sites abstaining. Their supplementary weight table gives exactly that — six patterns at +1, twelve at $-1/2$, nothing else — and the twelve are the three-base patterns of both orientations, so the abstentions are confirmed rather than assumed. Because the vote depends only on the site, and sites are i.i.d., the three topology scores are sums of i.i.d. three-vectors and are therefore *exactly* trivariate normal at every L , not merely asymptotically; the recovery probability is an exact bivariate orthant probability, computed by the same routine of §2.1 that maximum likelihood uses.

Zhang et al. (2025) compute the expected score difference — the margin — for a Farris and a Felsenstein zone tree, analytically, as a function of w_3 and of a branch-length ratio $1 \leq r < 12$ (their Fig. S1), and conclude that $-1/2$ “is the only weight x that leads to positive gap scores for both Felsenstein and Farris trees across all r values”. Their figure also states an identity, that $\text{Felsenstein}(-1/2, r) = \text{Farris}(-1/2, r)$ for all r . Our formulation reproduces that identity to machine precision, with residuals below 10^{-17} over $r = 1, \dots, 12$.

What the closed form adds is that the admissible set has a limit, and the limit is the value they chose. Farris bounds w_3 from below and Felsenstein from above, and the interval between the two bounds contracts:

$$\begin{array}{c|cccc} r & 4 & 8 & 12 & \geq 20 \\ \hline \text{admissible } w_3 & (-0.950, +0.500) & (-0.568, -0.385) & (-0.515, -0.480) & \{-0.500\} \end{array} \quad (12)$$

At $r = 20$ the window has closed to a single point, and at $r = 10^4$ the margin at $w_3 = -1/2$ is zero to machine precision in both zones (Fig. 7). So $-1/2$ is not a value that happens to work: it is the unique accumulation point of the admissible set as the branch-length ratio diverges, and their scan ends at $r = 12$ with the window 0.035 wide, at the last moment before it shuts. Their statement that “further decreasing the weights results in the wrong tree for Farris” is the lower bound moving up to meet the upper one.

That reading also explains the price. Their own words are that a negative vote is “chipping away at the score gap for Farris trees”, affordable “because Farris trees have a large gap based on two-nucleotide sites”. On the quartet used here the per-site margin is $+0.0877$ at $w_3 = 0$ and $+0.0012$ at $w_3 = -1/2$: a seventy-three-fold reduction, and one that §3.6.4 shows is not free.

The computation was checked against the binary rather than trusted. Running **caster-site** on simulated Farris quartets at five lengths, the closed form matches to a mean $|z|$ of 1.37, and at 1 Mbp it predicts 0.998 where the binary gives 0.983. Inverting the closed form at each length recovers the weight the binary behaves as if it used: -0.507 , -0.503 , -0.502 , -0.502 , -0.500 . The table is CASTER’s table.

3.6.3. The experiment

The pool is one 16-taxon Yule–Harding tree drawn once, before anything was measured, with 20 000 fragments of 500 bp simulated under it at twenty uniform rescalings of the branch lengths, so that topology is fixed and only tree length varies. I_{ss} is computed per fragment, and the fragments are split at the two measured cuts into CLEAN, MID and HIGH (Table 2). Four arms are compared — the three bands and a uniform random draw from the whole pool — each truncated to the same 1.025 Mbp, the largest length the smallest band supports. Length is therefore held constant and the arms differ only in composition. Three methods are run on each: IQ-TREE under JC, and both **caster-site** and **caster-pair**. Distances are Robinson–Foulds against the true tree, whose maximum on 16 taxa is 26.

The result at genome scale is negative (Fig. 8). From 10 kbp upward there are sixty condition-by-length-by-method combinations, and all but four recover the tree exactly; each of the four is off by a single internal branch ($\text{RF} = 2$). From 100 kbp upward every one of the thirty-six is exactly zero. Whatever the filter is doing, it is not doing anything that is visible in the answer.

The bands do separate below 5 kbp, and there the ordering is the one the argument predicts. MID — the data a 500 bp calibration rejects — performs like CLEAN and like the integrated arm from 10 kbp on, and recovers the tree exactly. It is real data that the index called saturated. HIGH, the band both calibrations reject, is where methods actually break: RF = 4 to 22 at 2 kbp. And even HIGH is recovered exactly by 5–10 kbp, which means that at these lengths no threshold among those tested locates data that is genuinely unusable.

Uncertainty, without replicates. Repeating the whole construction would be the obvious way to put error bars on those comparisons, and it is the wrong way, because the question is not how the result varies over independent studies but how it depends on which fragments were drawn. A delete-a-block jackknife answers that directly and costs a tenth of a replicate study. The fragments of each arm are split into ten consecutive blocks, each block is deleted in turn, and the arm is rebuilt without it, so every entry in the table is accompanied by ten leave-one-block-out values.

There are thirty-six computable cells — four arms, three lengths and three methods; the 500 bp cells are not computable, because at that length the alignment is a single fragment and there is no block to delete. Thirty-three of the thirty-six have a jackknife standard error of exactly zero. The three that do not are all at 10 kbp, and all three are the same event, one internal branch changing state: CLEAN with `caster-pair` and MID with `caster-site` have a leave-one-block-out mean of 0.200 against a full value of 0; HIGH with `caster-pair` has 1.800 against 2. At 100 kbp and at 1 Mbp, all twenty-four cells are exactly zero.

The reading is therefore not that the experiment is underpowered but that its comparisons do not depend on the sample: removing any tenth of the fragments leaves the answer where it was, everywhere except at the one length where a single split sits close enough to zero to be tipped by it. That is the expected shape and it is worth stating, because a jackknife which returned zero everywhere would be reported with more confidence than it earned.

Put together with §3.6.1, the picture is that the index identifies a real gradient — HIGH is harder than MID is harder than CLEAN — but places the boundary at a point where the gradient has barely begun, and then reports the placement to two decimals.

3.6.4. *The method that was supposed to be immune*

The discriminating prediction is about CASTER, and it does not survive contact with the data, in a specific and instructive way.

Zhang et al. (2025) establish that site weights with a negative vote on three-base sites give a statistically consistent estimator “regardless of varying rates of substitutions across the tree”, and their simulations show CASTER recovering Farris-zone species trees that concatenated maximum likelihood does not. The mechanism is exactly the one quantified in §3.6.2: the negative vote repairs the Felsenstein zone at a cost to the Farris margin.

On the HIGH band at 500 to 2000 bp, maximum likelihood does *better* than either CASTER version. At 2 kbp the heuristic search gives RF = 4 and a full search with model selection gives RF = 10, against 18 for `caster-site` and 22 for `caster-pair`; the weaker search is the one that flatters ML here, so the conservative number is the one quoted. There is no long-branch-attraction repair to be had in this setting, and the trade is therefore pure cost — the accounting of §3.6.2 showing up as accuracy.

This does not contradict the reported result, and the difference is the setting. Their Farris-zone comparison is run under the multispecies coalescent with recombination and 40% incomplete lineage sorting, where the concatenated likelihood method they compare against is inconsistent for reasons independent of long-branch repulsion; their failures there are obtained under both effects at once. The setting here has one tree, no coalescent and no ILS, which is the setting in which (3) applies and in which, as §3.5.1 shows, the difficulty is a shortage of data rather than an inconsistency. In that setting likelihood succeeds given enough sites, and the weighting has nothing to repair.

4. Discussion

4.1. What the criterion is

The index was introduced to answer one question — has this alignment accumulated too many substitutions for its tree to be recovered — and the answer it gives is a threshold comparison. What this paper shows is that the threshold is not a property of the substitution process. It is a property of the tree, through the gap D between that tree and its own best star approximation, and of the alignment length, through the noise; the two enter together in (3), and nothing else enters.

That reading dissolves the difficulty the literature has had in pinning the criterion down without denying that it works. A quantity defined by the motion of a boundary is a real quantity, and its gradient is real: the index separates configurations that differ in resolvability, and at genome scale it separates them in the right order. What it cannot do is carry a fixed cut across lengths, trees or taxon counts, because none of those is a saturating coordinate. The practical consequence is the one the last section measures: at the lengths genome-wide data are actually analysed at, the band a filter would remove is not distinguishable from the band it would keep.

4.2. Limits

- **Four taxa.** The closed form is quartet-specific. The finite-sample character of the length dependence is a property of any consistent method and should carry to larger trees, but the quantitative prediction does not. Quartet concordance factors are the natural extension, and the shuffle construction of §3.2 is a taxon-level operation that applies unchanged at any N .
- **JC69, no rate heterogeneity.** The factor sweep is one-dimensional in the tree geometry; base composition, GTR parameters and Γ rates are further factors and have not been varied.
- **Maximum likelihood, and therefore the long-branch-repulsion region.** D and v_2 as written are the maximum-likelihood quantities, so what is quantified throughout is the boundary of the region in which likelihood is the method that struggles. That region has a name: Siddall (1998) defined the Farris zone as the four-taxon case where the long branches are themselves sisters, parsimony recovers the correct tree and likelihood does not, and §3.5.1 shows the boundary in question to be exactly (3). Because that boundary is driven by D it moves with L , which is why Leuchtenberger et al. (2020)

describe the zone as “not unambiguously defined for ML inference”; (3) is the definition they declined to fix, and §3.4 is the boundary’s dependence on the tree.

- **On the older disagreement, and on where this paper stands in it.** Whether likelihood’s difficulty in that region should be called a bias at all was contested: Swofford et al. (2001) argued that parsimony’s success there follows from parsimony’s own bias happening to point the right way, and that likelihood is not actively repelled. The dispute has not closed; Zhang et al. (2025) report in 2025 that per-site likelihood scores “exhibited a strong bias against gene trees with adjacent short branches” in the Farris zone.

Within the setting of this paper the question is settled by construction rather than by argument, and it is worth being explicit about which setting that is. There is one tree, no coalescent, no recombination and no incomplete lineage sorting. In that setting the left side of (3) grows without bound in L , so recovery succeeds eventually and the failure is a shortage of data rather than an inconsistency — the quantitative form of Swofford et al. (2001)’s position, reached analytically. The two remedies then separate cleanly: switching to parsimony exploits a bias that happens to point the right way, whereas adding data removes the difficulty, and (3) says by how much.

That conclusion is not transferable to the setting Zhang et al. (2025) work in, for the reasons set out in the last item of this section.

- $I_{ss,c}$ is defined here by resampling at a 95% level, and the tabulated values of Xia et al. (2003) are not used numerically. The reason is not that their numbers are on a different scale; it is that they are numbers for different trees, and §3.4 shows the tree to be the dominant input. On the same alignment length, the two shapes used throughout this paper give $I_{ss,c}$ values of 0.33 and 0.99; a threshold is a property of a configuration before it is a property of a length, so agreement with a table drawn from other configurations would be a coincidence rather than a check.

What the comparison can use is the *shape* of their table, and it argues for the reading above rather than against it: their critical value is tabulated separately by alignment length, by number of taxa and by topology. A threshold that measured saturation would depend on none of the three, because saturation is a property of the substitution process alone; a threshold that measures resolution power depends on all three, and (3) says how — L directly, and the other two through D . The three-fold dependence of the table is therefore itself evidence for the power reading, and §3.5.1 accounts for the topology axis qualitatively by reproducing the ordering Xia et al. (2003) report.

- **The band experiment is one tree, one model, one setting.** §3.6 uses a single 16-taxon Yule–Harding tree held fixed while its branch lengths are rescaled, JC69 without rate heterogeneity, one gene tree with no coalescent and no recombination, and a four-taxon pattern of long branches that arises in the drawn tree rather than being constructed. The two cuts are measured at 50 replicates per point, which is enough to place the endpoints but not the intermediate lengths — the $I_{ss,c}$ estimates at 1000, 2000 and 5000 bp are not monotone in L and are not used. The uncertainty on the band comparisons comes from a delete-a-block jackknife over the fragment pool

rather than from independent replicates; it returns a standard error of zero wherever the answer is stable under removing a tenth of the data, and non-zero only where a statistic sits near a boundary. That is a statement about the sensitivity of the result to which fragments were drawn, which is the relevant variability here, but it is not a confidence interval over replicate studies.

- **One CASTER result is a single-tree result.** The finding in §3.6.4 that maximum likelihood outperforms both CASTER versions on saturated short alignments holds in the setting just described. It is not evidence against the consistency guarantee of Zhang et al. (2025), which is a statement about the multispecies coalescent, and the comparison there is against a concatenated likelihood method that is inconsistent under that model for reasons unrelated to long-branch repulsion. Nothing here bears on that case.

Data and code availability

All code and every result table are archived in two places. The repository at https://github.com/Gipsy-The-Shellier/Iss_simulations holds both papers of this series, and the tag `rs-preprint` is the version this manuscript was built from. The same material is deposited at Zenodo, <https://doi.org/10.5281/zenodo.23470551>. Nothing is behind a request.

The results come from the scripts in `Part_2/scripts/`, one per result set: `s27` (boundary machinery), `s32` (resampling definition of $I_{ss,c}$), `s34` (the I_{ss} -preserving path), `s35` (matched I_{ss}), `s36` (factor decomposition), `s37` (pointwise identity), `s39` (arrangement family), `s40` (domain and slicing), `s42` (the shared-origin identity), `s43` (the saturation control), `s44` (discard test), `s46` (fragment pool), `s47` and `s47b` ($I_{ss,c}$ against alignment length), `s48` (the band experiment and its block jackknife), `s49` and `s50` (the three-nucleotide weight and its validation against the CASTER binary), `s51` (the same matrices under ModelFinder), `s53` (the admissible-weight window), `s41` and `s52` (figures). No external data are used.

Author contributions

Z.-J.X. conceived the study, carried out the analysis, and wrote the manuscript.

Competing interests

The author declares no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethics statement

Not applicable. This study analyses published sequence alignments and simulated data; it involves no human participants, no human data, and no animals.

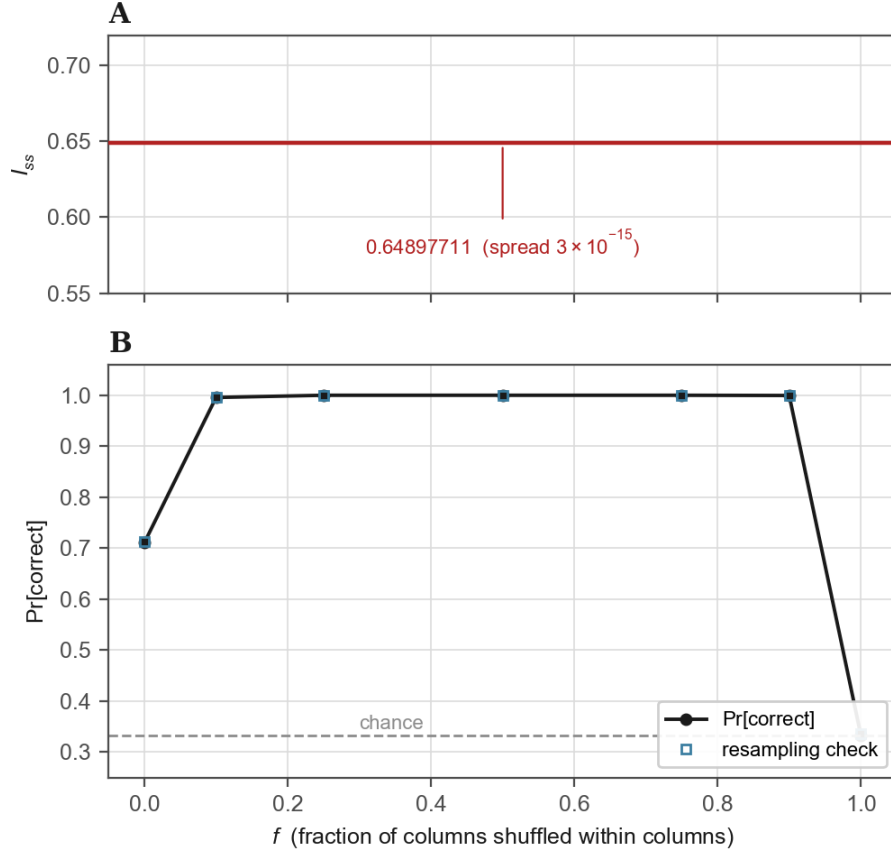


Figure 1: I_{ss} is exactly constant along a family of alignments whose recovery rate is not. Taxa are permuted within each column, independently per column, for a fraction f of columns. Panel A: the preservation is exact — $I_{ss} = 0.64897711$ at every f , spread 3×10^{-15} . Panel B: the same axis, and the recovery rate runs from 0.711 through 1.000 to the chance value $1/3$. Squares are an independent multinomial resampling check on the mixture distribution.

Use of AI-assisted tools

Claude Code (Anthropic), an AI coding assistant, was used in the preparation of this work. It assisted in drafting and editing the manuscript text, in writing and debugging the analysis code, and in deriving and checking the mathematical results reported here. All content was reviewed by the author, who takes full responsibility for the work and for its conclusions; the tool is not an author and cannot take responsibility for it.

4.3. Method notes

A fitted branch length is not evidence.. A projection onto the star is not established by the fitted internal branch being small. Near saturation the objective is flat along that branch, so an optimiser starting away from the star stops wherever its gradient runs out, having gained nothing; a small fitted branch length is a statement about where the optimiser stopped. Judged by objective instead, no configuration among the 219 scanned shows a gain over the star above 2.5×10^{-6} .

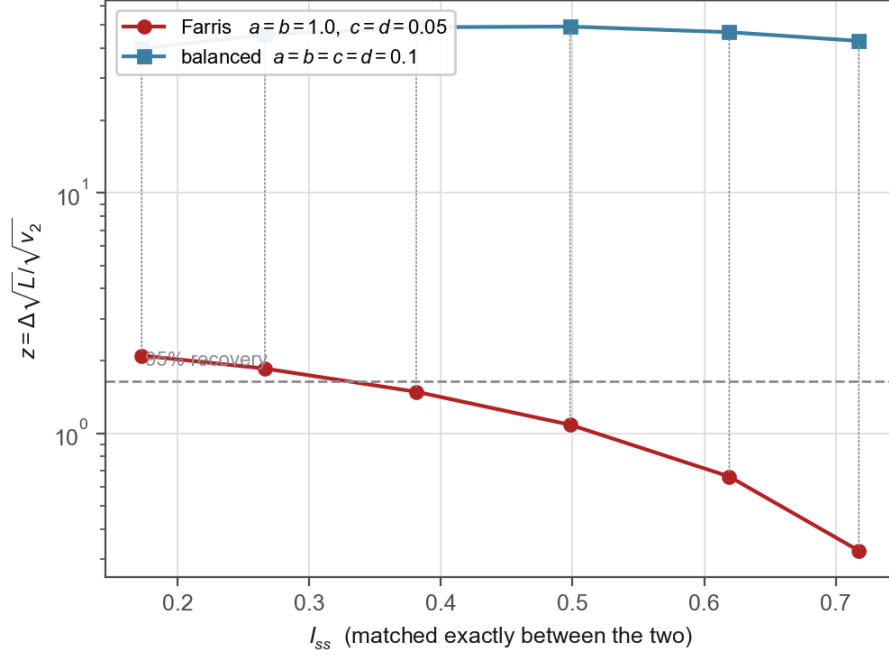


Figure 2: The same conclusion inside the tree family, without shuffling. For each I_{ss} value a Farris tree and a balanced tree are matched exactly on I_{ss} ; both are genuine trees growing under the same model. The standardised signal differs by factors of 19 to 125. At $I_{ss} = 0.1727$ the Farris tree’s internal branch is 0.003 against the balanced tree’s 0.096.

The degenerate corner.. Permuting to $f = 1$ collapses all three fits onto the same star, so $v_2 = v_3 = c_{23} \approx 0$ and $\rho = c_{23}/\sqrt{v_2 v_3}$ is $0/0$. Clipping ρ to 1 and taking the two-way branch returns $\Phi(0) = 1/2$ where the answer is $1/3$. The two-way reduction is valid only when the two *wrong* topologies coincide while the truth remains distinguishable; when all three coincide it is a genuine three-way tie. The construction of §3.2 is the only one in the paper that reaches this corner, which is as it should be: it is the only one whose purpose is to remove all signal.

Where the maximum-likelihood column comes from.. The maximum-likelihood entries of §3.6 are quoted from a full search with model selection wherever that search and the faster fixed-model search disagree. Model selection is expensive at the top of the length axis — on a 1 Mbp alignment the model set advanced at roughly one model per minute — so it was run on every arm up to 100 kbp and above that only where the two searches differ; where they agree, the faster search is reported. ModelFinder selects JC+G4 on every arm on which it was run, which is itself evidence about the construction: the band matrices concatenate fragments drawn at scales spanning two orders of magnitude, so the alignment is rate heterogeneous between fragments and a single-rate fit is misspecified. The model search finds exactly this, unprompted.

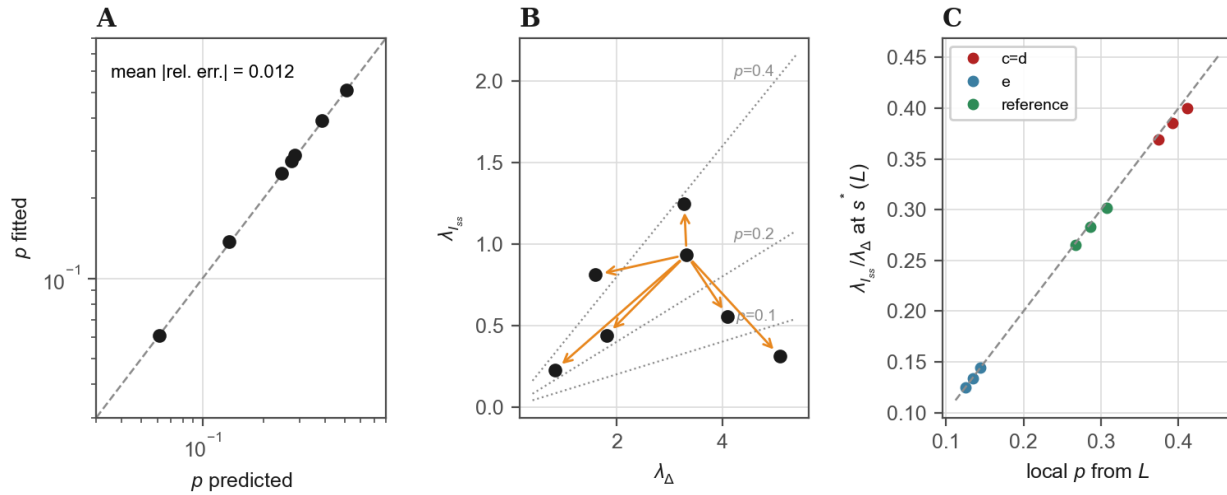


Figure 3: The exponent is a ratio of two measurable rates. Panel A: predicted against fitted p over six configurations, on the identity line; mean absolute relative error 1.4%. Panel B: the two rates as a plane, with iso- p lines; each arrow is one factor changed from the reference. Panel C: the pointwise comparison, local exponent from L against the locally measured rate ratio at the critical scale, for three configurations.

References

- Duchêne, D.A., Mather, N., Van Der Wal, C., Ho, S.Y.W., 2022. Excluding loci with substitution saturation improves inferences from phylogenomic data. *Systematic Biology* 71, 676–689. doi:10.1093/sysbio/syab075.
- Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* 27, 401–410. doi:10.2307/2412923.
- Hendy, M.D., Penny, D., 1989. A framework for the quantitative study of evolutionary trees. *Systematic Zoology* 38, 297–309. doi:10.2307/2992396.
- Jermiin, L.S., Misof, B., 2020. Measuring historical and compositional signals in phylogenetic data. doi:10.1101/2020.01.03.894097.
- Leuchtenberger, A.F., Crotty, S.M., Drucks, T., Schmidt, H.A., Burgstaller-Muehlbacher, S., von Haeseler, A., 2020. Distinguishing felsenstein zone from farris zone using neural networks. *Molecular Biology and Evolution* 37, 3632–3641. doi:10.1093/molbev/msaa164.
- Siddall, M.E., 1998. Success of parsimony in the four-taxon case: Long-branch repulsion by likelihood in the farris zone. *Cladistics* 14, 209–220. doi:10.1111/j.1096-0031.1998.tb00334.x.
- Strimmer, K., von Haeseler, A., 1997. Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment. *Proceedings of the National Academy of Sciences* 94, 6815–6819. doi:10.1073/pnas.94.13.6815.

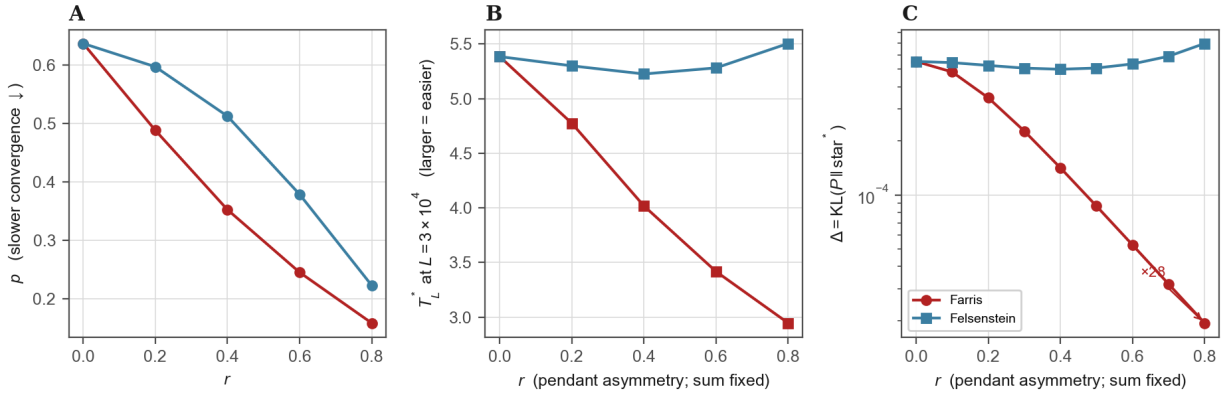


Figure 4: One quantity, three appearances. The pendant sum is fixed at 4, the internal branch at 0.2, and hence the total tree length at 4.2 throughout; only the arrangement moves. Panel C is the point: D is exactly the KL distance from the tree to its best star approximation, and the Farris arrangement drives it down by a factor of 28 while the Felsenstein arrangement leaves it flat. That is what the Farris zone is — the region of tree space nearest the star. Panels A and B are its two downstream consequences: the length exponent and the critical tree length. Over the same range I_{ss} moves by 2.6%.

Su, Z., Townsend, J.P., 2015. Utility of characters evolving at diverse rates of evolution to resolve quartet trees with unequal branch lengths: analytical predictions of long-branch effects. *BMC Evolutionary Biology* 15, 86. doi:10.1186/s12862-015-0364-7.

Swofford, D.L., Waddell, P.J., Huelsenbeck, J.P., Foster, P.G., Lewis, P.O., Rogers, J.S., 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Systematic Biology* 50, 525–539. doi:10.1080/106351501750435086.

Townsend, J.P., 2007. Profiling phylogenetic informativeness. *Systematic Biology* 56, 222–231. doi:10.1080/10635150701311362.

Townsend, J.P., Su, Z., Tekle, Y.I., 2012. Phylogenetic signal and noise: Predicting the power of a data set to resolve phylogeny. *Systematic Biology* 61, 835–849. doi:10.1093/sysbio/sys036.

Xia, X., Xie, Z., Salemi, M., Chen, L., Wang, Y., 2003. An index of substitution saturation and its application. *Molecular Phylogenetics and Evolution* 26, 1–7. doi:10.1016/S1055-7903(02)00326-3.

Xu, Z.J., 2026. Revisiting xia’s index of substitution saturation from a mathematical perspective. Preprint.

Zhang, C., Nielsen, R., Mirarab, S., 2025. Caster: Direct species tree inference from whole-genome alignments. *Science* 387, eadk9688. doi:10.1126/science.adk9688.

Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S., 2018. Astral-iii: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19, 153. doi:10.1186/s12859-018-2129-y.

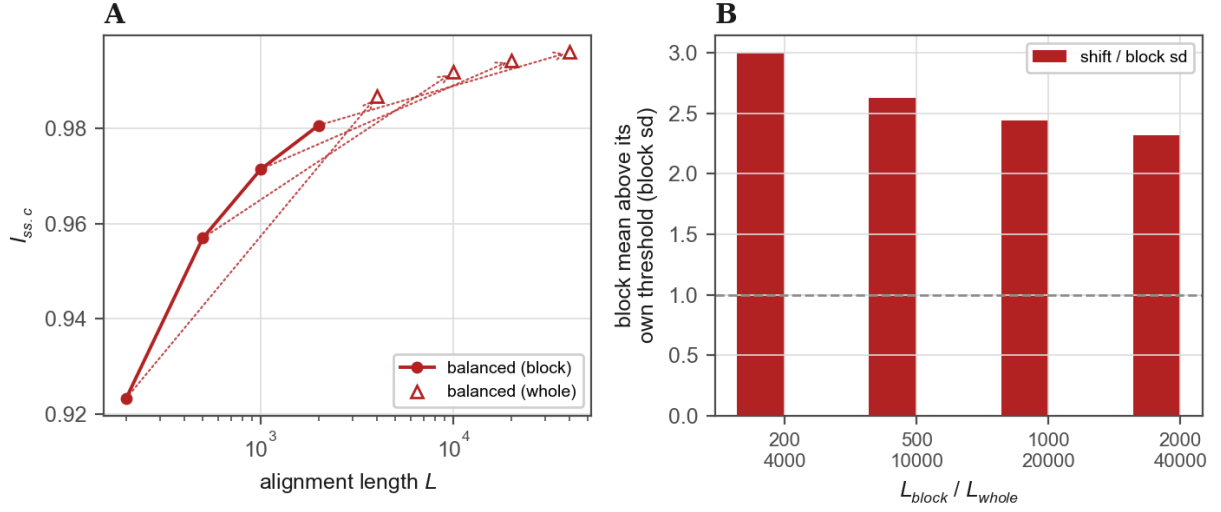


Figure 5: Where the criterion exists, and what slicing does to it. Panel A: $I_{ss,c}$ against alignment length, with arrows from the block's threshold to the whole's. Panel B: the block's expected I_{ss} is the whole's, since the blocks are slices of the same alignment, whereas the threshold they would be judged against is the one tabulated for the block's length. The block mean therefore sits 2.4 to 3.7 block standard deviations above its own threshold, and essentially every block is declared saturated while the whole passes.

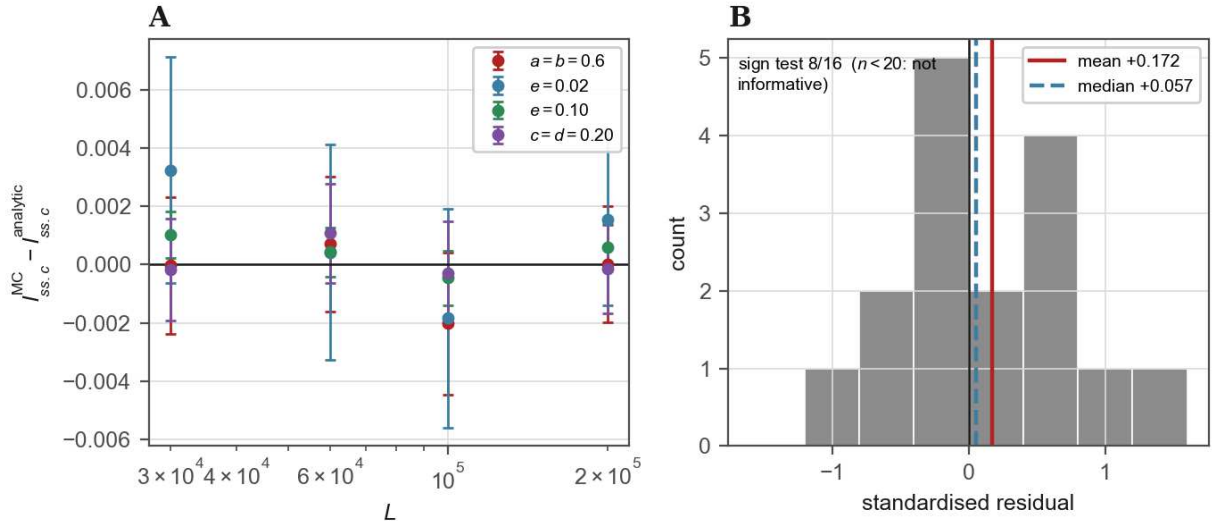


Figure 6: The analytic chain against our own resampling definition of $I_{ss,c}$, with no external data and no free parameters. Panel A: residuals with the per-point uncertainty, which varies by an order of magnitude across the sweep. Panel B: standardised residuals; the mean and median agree to within their uncertainty and the signs are split. With 16 points the sign test has no power, so this is a failure to detect a bias rather than evidence of none.

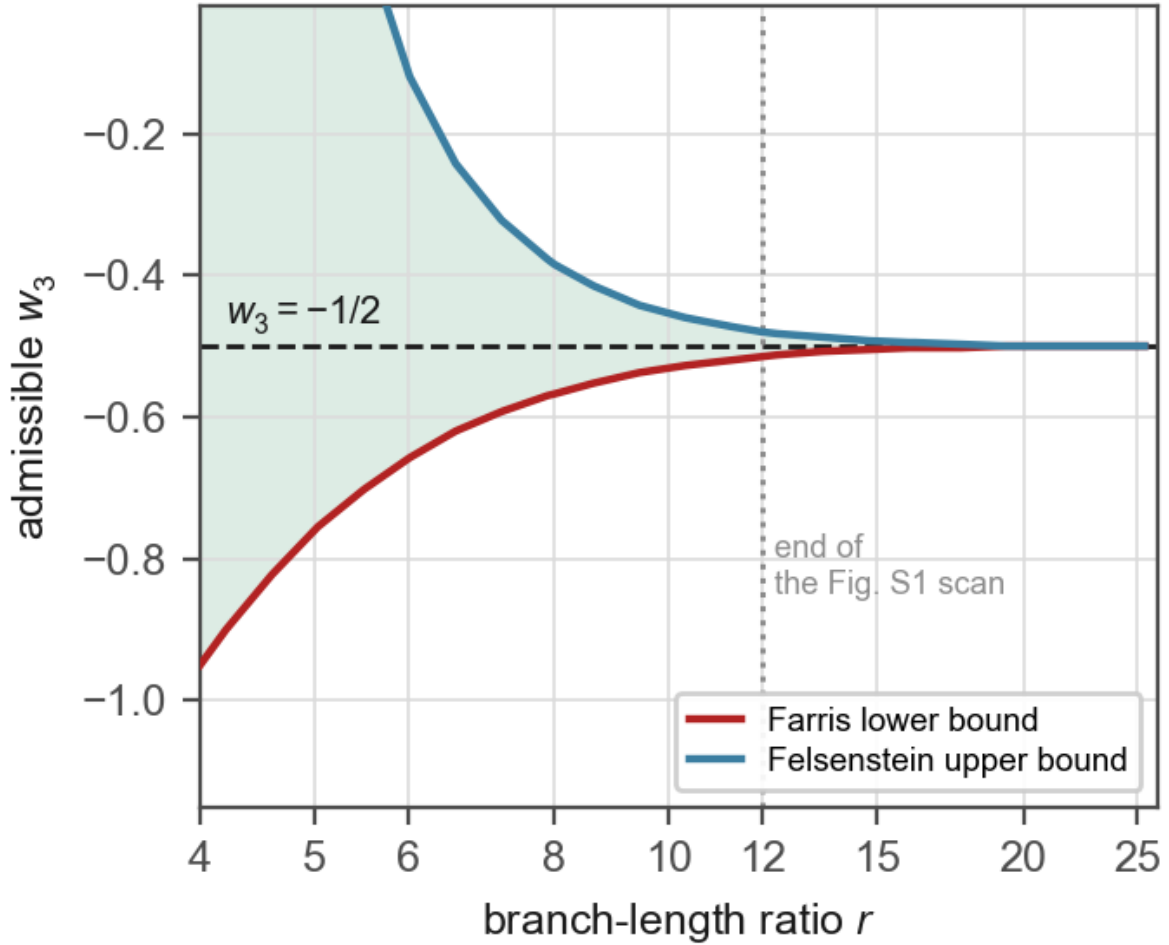


Figure 7: The admissible three-nucleotide weight closes onto $-1/2$. For each branch-length ratio r — long branches $0.1r$, short branches 0.1 , internal 0.1 , the parametrisation of the CASTER supplementary figure S1 — two constraints apply: the Farris-zone margin requires w_3 above the red curve and the Felsenstein-zone margin requires it below the blue one. The admissible set is the shaded region. It is the whole axis for $r < 2$, then closes monotonically from both sides, and by $r \approx 19$ it is the single point $-1/2$. The supplementary scan stops at $r = 12$ (dotted line), where the window is still 0.035 wide: the last moment before it shuts. $-1/2$ is therefore not a value that happens to work for both zones but the unique accumulation point of the admissible set.

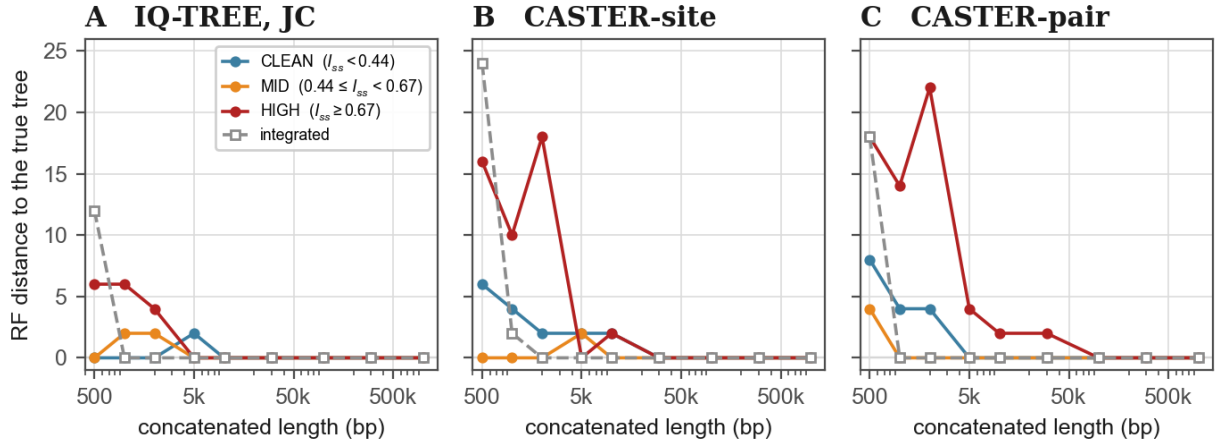


Figure 8: The band experiment. Fragments are banded by I_{ss} at the two cuts measured on the same tree, each arm is truncated to the same 1.025 Mbp, and Robinson–Foulds distance to the true tree is plotted against concatenated length for three methods. HIGH — the band both calibrations reject — is where methods break, and it breaks earlier and harder under CASTER than under maximum likelihood, which is the price of the weight computed in Fig. 7 paid where there is no long-branch attraction to repair. MID, the band a 500 bp calibration rejects and a 10 kbp calibration accepts, is indistinguishable from CLEAN and from the integrated arm at every length above 5 kbp. From 10 kbp upward every arm converges to exact recovery, so at the scale these data are actually analysed the band composition is not visible in the answer at all. The 500 bp and 1 kbp points are single fragments rather than concatenations and carry no rate.