Composite Reviewer Report

Manuscript: Chromosome number evolves at rates spanning seven orders of magnitude across eukaryotes

Copeland, McConnell, Barboza, et al.

Submitted to Science | Simulated Review Date: March 19, 2026

PART 1: SUMMARY & SIGNIFICANCE

This paper assembles 63,682 karyotypes across 55 eukaryotic clades and estimates dysploidy rates within a unified Bayesian framework (ChromePlus/diversitree). The central claim is that chromosome number evolutionary rates span seven orders of magnitude, that intraclade variance exceeds interclade variance, that avian chromosomal stasis is artifactual, and that life history rather than centromere architecture governs evolutionary tempo. The dataset is impressive and the question is important. However, the paper has serious methodological gaps: the MCMC chains are unusually short, time-calibration uncertainty is not propagated, the central variance claim lacks a formal statistical test, and the mechanism section substitutes narrative for quantitative evidence. As written, this is a strong pattern paper overselling itself as a mechanistic contribution.

Recommendation: Major Revision.

PART 2: MAJOR CONCERNS

1. The "seven orders of magnitude" claim

Problem: The abstract and main text repeatedly claim rates span "seven orders of magnitude, from fewer than 10^{-6} to nearly one event per million years." The range from 10^{-6} to 10^{0} is six orders of magnitude, not seven. For the claim to be correct, the minimum rate would need to be below 10^{-7} or the maximum above 10^{1}. The paper does not report actual minimum and maximum posterior median rates, making independent verification impossible.

Why it matters: This is the central quantitative finding and it appears in the title, abstract, and conclusion. A factual arithmetic error in the headline result would be deeply damaging to credibility at review and post-publication.

Fix: Report the exact minimum and maximum posterior median rates with credible intervals. If the range is six orders of magnitude, correct the title, abstract, and all instances in the text. If the range is truly seven, show the arithmetic explicitly.

Severity: Moderate (factual error, easily fixed by correcting the text or showing the arithmetic).

2. Cross-clade comparability and calibration heterogeneity

Problem: Rates are estimated on 55 phylogenies from heterogeneous sources (Smith & Brown megaphylogeny for plants, various clade-specific trees for animals, TimeTree for some groups). Trees differ in calibration methodology, fossil constraints, and molecular markers. The supplement states rates were estimated on unit-length trees and then divided by root age from the TimeTree of Life. This means every rate estimate is the ratio: posterior rate / root age. The root ages themselves come from different dating analyses with different priors and calibration points. No sensitivity analysis examines how calibration uncertainty propagates.

Why it matters: If one clade's root age is overestimated by 2x, its rate is underestimated by 2x. Root age uncertainties of 50–100% are not uncommon for deep nodes. Some of the seven-order-of-magnitude range could reflect calibration heterogeneity rather than biological rate variation. The paper acknowledges none of this.

Fix: Conduct a sensitivity analysis: shift all root ages by ±20% (or use published confidence intervals from TimeTree) and recompute rates. Report how the range changes. If the range shrinks from seven to five orders of magnitude under reasonable calibration uncertainty, the headline claim needs substantial revision.

Severity: Serious.

3. MCMC chain length

Problem: The methods state 1,000 MCMC iterations with 100 discarded as burn-in, yielding 900 post-burn-in samples. The justification is that "longer chains are unnecessary with this model that has relatively few parameters and all parameters had ESS values greater than 200." ESS > 200 is a necessary but not sufficient condition for convergence. With only 900 post-burn-in samples, ESS values near 200–300 (which several clades show in Table S2, e.g., Hemiptera ascending = 232, Primates descending = 244, Nothobranchiidae polyploidy = 222) suggest substantial autocorrelation in those chains. More critically, 1,000 steps cannot diagnose multimodality. No mention is made of running multiple independent chains or inspecting trace plots.

Why it matters: For complex state spaces (e.g., Orchidaceae with 1,904 species and potentially 40+ chromosome states, Pteridophyta with 1,465 species), 1,000 steps is orders of magnitude shorter than standard practice. If the posterior is multimodal, 1,000 steps will lock into one mode and produce misleadingly narrow credible intervals. The paper cannot claim its uncertainty estimates are trustworthy without ruling this out.

Fix: Run at least three independent chains per clade of at least 10,000 steps. Report Gelman-Rubin diagnostics. Show trace plots for representative clades spanning the complexity range (e.g., Cetacea with 34 species, Orchidaceae with 1,904). If results are unchanged, this strengthens the paper. If results change, the current analysis is unreliable.

Severity: Serious.

4. The intraclade > interclade variance claim

Problem: The abstract and main text assert that "intraclade variance exceeds interclade differences by more than an order of magnitude." This is a central quantitative claim that appears without any supporting statistical test, variance decomposition, or even a table of the relevant values. The text does not define what "intraclade variance" means operationally: is it the variance of posterior samples within a single clade? The variance of lineage-specific rate estimates within a clade (which the model does not produce, since ChromePlus estimates a single rate per clade)? The variance of chromosome numbers at the tips?

Why it matters: If "intraclade variance" refers to posterior uncertainty within each clade's rate estimate, then the claim is that uncertainty is larger than signal — which would undermine rather than support the paper's conclusions. If it refers to something else, the definition is absent. Either way, this claim needs a formal test: a nested ANOVA or hierarchical model partitioning rate variation into within-kingdom vs. within-clade components, conducted on log-rates.

Fix: Define the claim precisely. Conduct a variance decomposition (e.g., nested ANOVA on log-transformed median rates, or a hierarchical Bayesian model with kingdom/clade random effects). Report the actual variance components with confidence intervals.

Severity: Serious.

5. Mechanism section: pattern vs. process

Problem: The section "The Determinants of Tempo" presents compelling verbal arguments about why orchids evolve fast (pollinium reproduction, self-compatibility, fragmented habitats, small Ne) and odonates evolve slowly (large Ne, extensive gene flow, obligate outcrossing). However, not a single formal test of these predictors is presented. The one quantitative test in the paper — PGLS of genome size vs. dysploidy rate — is null ($p = 0.41$). The paper effectively demonstrates that centromere type does not predict rate (citing Ruckman et al. 2020), then substitutes a narrative about life history without testing it.

Why it matters: Science expects papers to advance mechanistic understanding beyond narrative. The orchid/odonate comparison is an N = 2 anecdote dressed up as evidence. Without formal tests of generation time, body mass, Ne proxies, mating system, or dispersal ability across the 55 clades, the mechanism section is speculation. The paper is being positioned as demonstrating that "life history and population structure rather than deep phylogenetic constraints governs the tempo of karyotypic change" (abstract), but it does not test this claim.

Fix: Either (a) assemble life-history trait data (generation time, body mass, range size/Ne proxy, mating system) for the 55 clades and run PGLS models predicting dysploidy rate, or (b) explicitly reframe the mechanism section as hypothesis-generating and remove causal language from the abstract and conclusion. Option (a) would make this a much stronger paper. Option (b) would be honest but may reduce the paper's suitability for Science.

Severity: Serious.

6. The bird microchromosome claim

Problem: The paper claims avian stasis is "largely an artifact of methodological resolution" because "historical surveys were frequently blind to the dynamics of microchromosomes." It then states that "by incorporating all chromosome transitions, including microchromosomes, into a unified phylogenetic framework, we find that every avian order in our dataset exhibits dysploidy rates that exceed the global background median." The supplement shows that the avian karyotype data come from Alfieri et al. (2024, bioRxiv). The paper does not specify whether these data include resolved microchromosome counts for each species or whether they represent total 2n counts that bundle microchromosomes into a single number.

Why it matters: If the karyotype data for birds already include microchromosome counts, then the claim about historical methodological blindness is valid — but the paper needs to explicitly state what data it has that previous studies lacked. If microchromosome counts are still imputed, inferred, or represent total 2n counts (as is common in avian cytogenetics), then the paper is claiming to resolve microchromosome dynamics without actually having microchromosome-level data, which would be circular.

Fix: State explicitly: for the avian taxa in this dataset, how many have independently resolved microchromosome counts vs. total 2n counts? What is the source of microchromosome resolution? If the data are total 2n counts (which include microchromosome variation implicitly), reframe the claim: the data still capture microchromosome dynamics, but the "resolution" framing suggests a methodological advance that may not exist in the data.

Severity: Moderate (likely fixable with reframing, but critical for the narrative).

7. Model adequacy

Problem: The main text states models "typically adequately captured" observed data. Table S4 reveals that 8 of 56 clades (including Magnoliaceae, which was excluded) failed posterior predictive checks on variance, and 8 failed on entropy. Several clades fail on both (Fabaceae, Lepidoptera, Rubiaceae, Scorpiones, Siluriformes). Additionally, 12 clades show prior-posterior overlap exceeding 40%, meaning the data have minimal influence on the posterior for these clades (Drosophilidae 44%, Orthoptera 57%, Tenebrionidae 57%, Hydrophilidae 63%, Marsupialia 68%, Curculionidae 68%, Coccinellidae 53%, etc.).

Why it matters: Clades where the model fails PPS checks are clades where the estimated rate may be a poor summary of the actual evolutionary process. Clades with high prior-posterior overlap are clades where the rate estimate is driven primarily by the prior, not the data. Including these clades in cross-clade rate comparisons and in the seven-orders-of-magnitude range without flagging this is misleading. The word "typically" in the main text obscures a 14% failure rate on variance and a 14% failure rate on entropy.

Fix: Report the exact number of clades failing PPS checks in the main text. Either exclude failed clades from the primary rate comparison or demonstrate that their inclusion does not change the headline findings. For clades with >50% prior-posterior overlap, discuss whether these rate estimates are informative or merely reflect the prior.

Severity: Moderate.

8. Within-clade rate homogeneity assumption

Problem: ChromePlus estimates a single set of rate parameters (ascending, descending, polyploidy, demiploidy) per clade, assuming rate homogeneity across the entire phylogeny within each clade. The paper's own central finding is that rates are highly heterogeneous. If rates vary substantially within a clade (which they almost certainly do — consider Lepidoptera with 322 species spanning Papilionoidea to micromoths), the single-rate estimate is an average that may not represent any actual lineage.

Why it matters: The cross-clade comparison is comparing averages of potentially wildly heterogeneous within-clade rates. The variance of these averages is not the variance of lineage-specific rates. The paper claims intraclade variance exceeds interclade variance, but the model is structurally incapable of estimating intraclade rate variance. This is a fundamental tension between the model and the claims.

Fix: Acknowledge this limitation explicitly. For a few focal clades, consider running analyses on subclades to estimate within-clade rate heterogeneity empirically. Alternatively, use a model that allows rate heterogeneity (e.g., hidden-rates Mk). At minimum, discuss how averaging over within-clade heterogeneity affects interpretation of the cross-clade comparisons.

Severity: Moderate.

9. Time calibration method and error propagation

Problem: The supplement states: "reported rates have been transformed into units of millions of years by dividing the posterior rate estimate by the tree depth. Median divergence estimates were obtained from the Time Tree of Life for taxa that subtended the root of the tree." This means the final rate for each clade is: (posterior rate on unit tree) / (single point estimate of root

age). No uncertainty in root age is propagated. The credible intervals on rates therefore reflect only MCMC uncertainty in the rate parameter, not calibration uncertainty.

Why it matters: TimeTree root ages often have confidence intervals spanning 20–50% of the point estimate. For Orchidaceae (root age 17 Ma), a plausible range might be 12–25 Ma. Dividing by 12 vs. 25 changes the rate by more than 2x. For Bryophyta (488 Ma), the absolute uncertainty may be 100+ Ma. Without propagating this, credible intervals are too narrow and the spread across clades is inflated. The claimed seven-order range may partly reflect calibration noise.

Fix: Obtain confidence intervals for root ages from TimeTree (or published sources) and propagate them through the rate calculation. Report rates as distributions that incorporate both MCMC uncertainty and calibration uncertainty. If TimeTree CIs are unavailable, apply a uniform ±20% perturbation as a minimum sensitivity analysis.

Severity: Serious.

10. AI-augmented data collection

Problem: The paper uses ChatGPT-5 for literature discovery, stating that "all LLM-derived outputs were treated strictly as leads and were subject to verification." However, verification can only catch false positives (wrong papers, wrong counts). It cannot catch false negatives (missed papers, missed counts). The paper reports no assessment of the false-negative rate of the LLM search process.

Why it matters: If the LLM systematically missed certain taxa (e.g., species described in non-English journals, older cytogenetic literature, gray literature), this could bias the dataset taxonomically. For a paper claiming global coverage, systematic gaps in discovery are concerning.

Fix: For 3–5 focal clades, compare the LLM-discovered dataset against a traditional comprehensive search (e.g., existing databases like the Chromosome Counts Database, Tree of Sex, Coleoptera Karyotype Database). Report how many records the LLM found that were in the database, how many it missed, and whether missingness correlates with any taxonomic or geographic variable. If coverage is >90% for these benchmark clades, the concern is mitigated.

Severity: Moderate (likely addressable given that the paper already uses established databases for many clades, as evident from Table S1).

PART 3: MINOR CONCERNS

1. The reference list mixes two Paperpile libraries (WuICVL and vM5f0W), suggesting incomplete bibliography management. References 30–37 appear to use a different library key than references 1–29.

2. Missing citations: Escudero et al. (2014, New Phytologist) on dysploidy in Carex; Carta et al. (2020) on chromosome number evolution in angiosperms; Guerrero & Kirkpatrick (2014) on fixation of chromosomal rearrangements; Lande (1979) on the effective population size theory of chromosomal speciation. The mechanism section would benefit from engaging with the theoretical population genetics literature on rearrangement fixation.

3. Figure 3 is described as a "conceptual model" and was generated by AI (ChatGPT-5). For Science, conceptual figures should be grounded in quantitative data. A figure plotting actual estimated rates against available life-history covariates would be far more informative than a cartoon, however attractive.

4. The Orchidaceae root age is listed as 17 Ma in the supplement. This seems very young for a family with Cretaceous pollen fossils. If the tree is from Smith & Brown (2018), confirm this root age is correct. An underestimated root age would inflate the rate, making orchids artificially the fastest-evolving clade.

5. Several clades have very low overlap between karyotype data and phylogeny tips (Phasmatodea: 11 species, Coccinellidae: 24, Hydrophilidae: 21, Blattodea: 30). These sample sizes push the limits of what Mk-type models can reliably estimate. The main text should acknowledge that rate estimates for these small-overlap clades carry substantially greater uncertainty.

6. The supplement reports that Brassicaceae has 45% unresolved polytomies and Fabaceae 56%. The tree resolution sensitivity analysis (Figs. S61–S62) shows that Orchidaceae, Solanaceae, and Rubiaceae show "pronounced changes" with "limited overlap" between original and rerun posteriors. This is a red flag that tree quality substantially affects rate inference for some plant clades, yet the main text does not flag this.

7. The term "demiploidy" is used throughout but never defined in the main text. Define it on first use.

8. The text alternates between "56 clades" and "55 clades" without always clarifying that Magnoliaceae was excluded. Standardize.

9. Supplement Table S1 lists some clades under "Various authors" without naming the specific student leader. For a CURE paper, attribution matters.

10. The prior-posterior overlap metric (Table S4) uses 70% as the flagging threshold, but 12 clades exceed 40% overlap. The choice of 70% is very permissive. Justify this threshold or use a more standard criterion.

11. The Odonata ESS values are anomalously high (46,954–143,821), suggesting the posterior is essentially flat and the chain is sampling the prior. The Odonata rate estimate should be treated with extreme caution and this should be discussed.

12. The paper does not cite chromEvol (Mayrose et al. 2010, Glick & Mayrose 2014), which is the most widely used tool for chromosome number evolution. A comparison or discussion of why ChromePlus was preferred over chromEvol is warranted.

PART 4: QUESTIONS FOR THE AUTHORS

1. What is the actual minimum and maximum posterior median dysploidy rate across your 55 clades, in events per million years, and which clades hold these extremes?

2. How many of your avian karyotype records include independently resolved microchromosome counts versus total 2n counts, and what is the primary source of those counts?

3. Were multiple independent MCMC chains run for any clade? If so, were Gelman-Rubin convergence diagnostics computed? If not, what is the basis for claiming convergence beyond ESS?

4. How is "intraclade variance" defined operationally? Is it posterior uncertainty within a clade, variance of tip chromosome numbers within a clade, or something else? On what scale (raw or log) was the comparison to interclade variance computed?

5. For the 12 clades with prior-posterior overlap exceeding 40%, do you consider the rate estimates informative? If Curculionidae (68% overlap) and Marsupialia (68% overlap) are included in cross-clade comparisons, how do you distinguish signal from prior?

6. The Orchidaceae root age of 17 Ma is substantially younger than estimates in some recent analyses (e.g., Givnish et al. 2015 place crown Orchidaceae at ~90 Ma). Which node does your 17 Ma represent, and how does this affect the rate estimate?

7. Were any formal tests of life-history predictors (generation time, body mass, Ne proxies, mating system, range size) conducted beyond the genome size PGLS? If so, why are they not reported? If not, on what basis do you claim life history governs rate?

8. For clades where posterior predictive simulations failed (e.g., Fabaceae fails both variance and entropy), what does this imply about the reliability of the rate estimate for those clades?

9. What is the expected effect of using point estimates of root age (from TimeTree) rather than propagating calibration uncertainty? Have you evaluated how your rate distributions would change if root ages were perturbed by their published confidence intervals?

10. Several plant phylogenies contain >40% unresolved polytomies. Your sensitivity analysis shows "pronounced changes" for Orchidaceae, Solanaceae, and Rubiaceae after pruning polytomy-descended tips. Given that Orchidaceae is the headline fastest-evolving clade, how robust is this result to tree quality?

PART 5: OVERALL ASSESSMENT

This is an ambitious and important paper with a genuinely impressive dataset. The question is timely, the taxonomic scope is unprecedented, and the CURE framework is a model for inclusive science. However, the paper currently has too many unsupported quantitative claims and too little methodological rigor for Science. The MCMC chain length is indefensible by modern standards, the time-calibration uncertainty is completely ignored, and the mechanism section is narrative without quantitative support. If the authors run longer chains, propagate calibration uncertainty, formally test life-history predictors, and tighten the claims to match the evidence, this could be a landmark paper in Science or Nature. As written, it is better suited to Nature Ecology & Evolution or PNAS, where the broader framing can be preserved but the evidentiary bar is slightly lower. The single revision that would most increase its chances at Science: replace the narrative mechanism section with a formal phylogenetic regression of life-history traits on dysploidy rate across clades.