Research Square

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

Direct whole-plastome alignment systematically inflates long-branch artifacts in chloroplast phylogenomics: a 468-paper audit and three-method reanalysis

Zhi-Jie Xu

zjxmolls@outlook.com

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

Research Article

Keywords: chloroplast genome, plastid phylogenomics, long-branch attraction, multiple sequence alignment, SibeliaZ, FreeRate, single-copy orthologs, reproducibility

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11584316/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.

Direct whole-plastome alignment systematically inflates long-branch artifacts in chloroplast phylogenomics: a 468-paper audit and three-method reanalysis

Zhi-Jie Xu[1]

[1]**College of Biological Science and Medical Engineering, Donghua University, Shanghai, China**

Corresponding author:
Zhi-Jie Xu[1]

Email address: zjxmolls@outlook.com

ABSTRACT

Plastid genomes are the workhorse of plant phylogenomics, but their quadripartite structure, inverted-repeat (IR) and small single-copy (SSC) orientation changes, gene rearrangements and length-variable noncoding regions make direct whole-plastome multiple sequence alignment (MSA) problematic. We audited 468 chloroplast/plastid genome papers published in *Mitochondrial DNA Part B* between October 2023 and October 2026, extracted their materials and methods, and flagged 53 papers with suspicious long branches or topology. Each candidate was reanalyzed with three independent strategies: (C) whole-genome MAFFT + trimAl + IQ-TREE, mimicking the common published workflow; (A) SibeliaZ locally collinear blocks (LCBs) concatenated and analyzed with a patched RAxML-NG implementing T92+F+FreeRate (2–10 categories, BIC-selected); and (B) single-copy orthologs + IQ-TREE TEST-MERGE. We compared Robinson–Foulds distances and branch-length diagnostics across methods and manually reviewed every case against the published tree. Thirty-six papers were strict positives: the published long branch was reproduced by C and disappeared or shortened under A/B, giving a confirmed positivity rate of 7.69% (95% CI 5.61–10.47%; 10.75% among 335 branch-length-auditable papers); 34 of the 36 (94.4%) had used direct whole-plastome MSA in the original materials and methods, including one case with an ambiguous figure legend that we mark with an asterisk and test separately. The remaining candidates comprised eight CDS/PCG-based studies, six whole-genome cases attributable to local-clock or genuine rate heterogeneity, and three negatives, one of which was reclassified after re-review. A BLAST screen of the 24 positives that lacked SSC/IR assays, plus a LASTAL re-check of the nine whose original assays were negative, attributed 16 additional cases to SSC orientation discordance and 16 to complex rearrangement; all 36 positives now have a structural attribution. Positivity was 20.0% in 2023 and approximately 6.7% in 2024–2026. All 36 positives deposited supplementary material, but none deposited a machine-readable alignment, partition scheme or tree topology. Direct whole-plastome MSA is therefore not a neutral shortcut: in this corpus it systematically inflates long branches, and distance-matrix residuals localize the effect to the curated long-branch taxa (26/27 C–A and 25/27 C–B comparisons; A–B shows no directional difference). Shared alignment errors can also produce supported clades when several close relatives carry the same structural variant, so topology alone is not a safeguard. It is not the only cause of long branches; routine locus extraction, LCB- or ortholog-aware alignment, alignment inspection, and explicit rooting/IR–SSC checks are required to separate alignment artifacts from biological rate variation.

Keywords: chloroplast genome; plastid phylogenomics; long-branch attraction; multiple sequence alignment; SibeliaZ; FreeRate; single-copy orthologs; reproducibility

INTRODUCTION

Plastid genomes are small, structurally conserved and inexpensive to assemble, and they have become the default source of data for plant phylogenomics (Gitzendanner et al., 2018; Jin et al., 2020; Tillich et al., 2017). The standard workflow is well established: assemble and annotate the plastome, extract protein-coding genes (PCGs) or other loci, align each locus, concatenate the loci with an explicit partition scheme, and infer a tree under maximum likelihood or Bayesian inference (Dong et al., 2022; Meckoni et al., 2023; Zhang et al., 2020, 2025). This workflow is implemented in organelle-specific platforms such as PhyloSuite and PlastidHub (Zhang et al., 2020, 2025), in plastome-specific automated pipelines such as PAPAplastomes (Meckoni et al., 2023), and in the widely used gene-matrix practice of plastid phylogenomics (Dong et al., 2022; Gitzendanner et al., 2018).

The structural biology of the plastome complicates this otherwise simple pipeline. Most plastomes have a quadripartite organization comprising a large single-copy region (LSC), a small single-copy region (SSC) and two inverted repeats (IRs). In short-read assemblies the two IR copies are identical or nearly identical, so reads map equally well to both copies and the assembly collapses to a circular LSC–IR–SSC–IR path. That path can be split into two alternative isomers that differ in SSC orientation; short reads alone cannot distinguish them, and the choice is rarely reported. Long-read or PCR/Sanger evidence is needed to resolve the orientation unambiguously (Ferrarini et al., 2013; Twyford and Ness, 2017). Gene order can also rearrange, and noncoding regions can differ substantially in length (Wicke et al., 2011).

A direct whole-plastome MSA must therefore align sequence that is not homologous across the whole matrix, and gap placement becomes arbitrary. Two additional sources of structural non-homology are common. First, DNA moves among the nuclear, plastid and mitochondrial compartments, so plastome assemblies can contain nuclear or mitochondrial insertions and nuclear assemblies can contain plastid-derived sequences (Timmis et al., 2004). Second, cpDNA itself rearranges through inversions, translocations and gene duplications, changing gene order and orientation (Wicke et al., 2011). A direct whole-plastome MSA treats these events as indels or mismatches rather than as structural changes, so the alignment can mix non-homologous positions and generate spurious branch lengths. The problem is not merely computational. Roestel et al. showed that aligning complete plastomes by algorithm alone is less reliable than partitioning the plastome into coding and noncoding regions and inspecting the alignments on conserved sequence motifs; different plastome partitions can yield significantly different trees (Roestel et al., 2024).

The protocols that are actually used for organelle phylogenomics avoid direct whole-genome alignment. The Bio-protocol pipeline for mitogenome phylogenomics extracts each protein-coding gene and rRNA separately, aligns PCGs with MACSE and rRNAs with MAFFT, and then concatenates the gene alignments; the authors state explicitly that aligning genes separately, rather than aligning full mitogenomes, makes it easier to clean alignments, check reading frames and remove stop codons (Winn et al., 2025). PhyloSuite and PAPAplastomes follow the same logic for organelle and plastome data (Meckoni et al., 2023; Zhang et al., 2020), and LCB- or ortholog-based tools such as HomBlocks, SibeliaZ and ORPA were developed precisely because whole-genome MSA is not robust to rearrangements (Bi et al., 2018, 2024; Minkin and Medvedev, 2020). Direct whole-plastome MSA is thus not a neutral default; it is a methodological shortcut that departs from the standard protocol.

What has been missing is a quantitative estimate of how often this shortcut changes a published conclusion. Case studies and methodological reviews warn about alignment problems and long-branch attraction (Bergsten, 2005; Felsenstein, 1978; Roestel et al., 2024), but they do not estimate prevalence across a journal corpus, and they rarely compare whole-genome, LCB and ortholog-based analyses on the same taxon sets. We therefore audited three years of chloroplast/plastid papers in *Mitochondrial DNA Part B*, a fast-turnover taxonomic journal with a large chloroplast-genome output, and reanalyzed every suspicious case with three independent methods. Our aims were to (i) quantify the prevalence of strict long-branch positives, (ii) identify which methodological choices predict artifacts, and (iii) provide a reproducible framework and practical reporting standard for plastid phylogenomics.

MATERIALS AND METHODS

Corpus and screening

We queried Europe PMC for chloroplast and plastid genome papers published in *Mitochondrial DNA Part B* between 24 October 2023 and 24 October 2026. The corpus contained 468 records, of which 355 had a full-text tree figure; 335 were branch-length-auditable and 21 were cladogram-only (20 matched to the index; 2449725 was not present). Records were excluded if they concerned mitochondrial or nuclear genomes, were too short to contain a phylogenetic analysis, or lacked a complete plastome. For every record we extracted the materials and methods: locus choice (whole plastome, PCGs, CDSs or individual loci), multicopy handling, alignment software and strategy, trimming, tree-building method, substitution model, partitioning, software, outgroup, and any use of general-purpose desktop software such as MEGA for genome-wide tree inference without explicit partitioning or model selection (Kumar et al., 2024).

A paper was flagged as a candidate if its published tree contained a long terminal branch, a suspicious clade placement, or a topology inconsistent with closely related taxa. The screen was intentionally sensitive: 53 papers (11.3% of all records; 14.9% of tree-bearing records; 15.8% of branch-length-auditable records) were carried forward. All 53 were manually reviewed and assigned a final disposition.

Three reanalysis strategies

For each candidate we reconstructed the taxon set from the paper's accession list and, where necessary, recovered missing CDS annotations with a conservative tblastn procedure (single high-confidence HSP, $\geq 80\%$ query coverage, start methionine, no internal stop, $\geq 80\%$ of reference length; multi-HSP or ambiguous models rejected).

Method C (whole-genome). Sequences were aligned with MAFFT (Katoh and Standley, 2013), trimmed with trimAl in automated1 mode (Capella-Gutierrez et al., 2009), and analyzed with IQ-TREE 3 using ModelFinder and 1,000 ultrafast bootstrap replicates (Kalyaanamoorthy et al., 2017; Wong et al., 2026). Method C intentionally mimics the common published workflow.

Method A (LCB + T92 + FreeRate). Locally collinear blocks were identified with SibeliaZ (Minkin and Medvedev, 2020); each block was aligned with MAFFT, concatenated without partitions, and analyzed with a patched build of RAxML-NG v1.2.2 that implements T92 natively (Kozlov et al., 2019). T92 is used because LCB orientation is arbitrary: its exchangeabilities are K2P-like and its equilibrium frequencies satisfy $\pi_A = \pi_T$ and $\pi_C = \pi_G$, so the model is invariant under reverse complementation. Because neither IQ-TREE 3 nor stock RAxML-NG provides a native T92 model with ML-estimated parity-constrained frequencies, we added T92 to the RAxML-NG DNA model table and implemented a one-dimensional Brent optimizer for the single GC parameter; both the GC content and the transition/transversion ratio are estimated by maximum likelihood. The number of FreeRate categories was selected from 2 to 10 by BIC on the same concatenated alignment with IQ-TREE 3 (Kalyaanamoorthy et al., 2017; Wong et al., 2026); the selected count was then used for a single-start RAxML-NG SPR search; four categories were selected for all 38 alignments. The patched binary and the source patch are provided with the manuscript.

Method B (single-copy orthologs). OrthoFinder was used to identify orthogroups (Emms and Kelly, 2019); each orthogroup was aligned with MAFFT, concatenated into a partitioned matrix, and analyzed with IQ-TREE 3 TESTMERGE (Wong et al., 2026). When OrthoFinder failed on Windows, we used a name-based single-copy matrix (one representative per gene name per taxon; missing genes allowed) and labeled the result as a fallback.

Trees were rooted on the paper's stated outgroup where available; otherwise we used the outgroup employed in the original analysis and recorded any root correction.

Branch-length and topology diagnostics

We calculated normalized root-to-tip distance and terminal branch length divided by tree height for each taxon, and flagged outliers with robust z-scores. We compared C, A and B using Robinson–Foulds distances (Robinson and Foulds, 1981) and inspected clade-level branch lengths. Because no sampling dates were available, the root-to-tip plots are strict-clock diagnostics rather than temporal regression; the logic follows TempEst-style root-to-tip exploration (Rambaut et al., 2016). To localize branch-length differences to individual taxa, we also analyzed patristic distance matrices. For each case and method pair (C–A, C–B, A–B), distances among shared taxa were normalized by the median off-diagonal distance and regressed on the log scale: $\log(d_{\text{second}}) = a + b\log(d_{\text{first}})$. Residuals were assigned to both taxa in each pair, and median residuals per taxon were compared between the curated long-branch taxa and all other

taxa. Because pairwise distances are non-independent, inference was performed at the dataset level with sign tests across cases, not at the level of individual pairs. Normalization by a per-matrix constant leaves residuals and slope unchanged in a log–log OLS with intercept; it makes distances comparable across methods for display.

BLAST re-screen for SSC orientation discordance and rearrangement

For every positive that lacked an SSC/IR assay in the original paper, we selected a normal non-suspect ingroup as the accession with the strongest BLAST hit to the suspect sequence, then compared the suspect to that comparator with BLASTN 2.17.0+ (Camacho et al., 2009) (`-task blastn`, $E = 10^{-10}$, dust disabled). We recorded all high-scoring segment pairs and searched for query intervals covered by reverse-strand hits but not by any forward-strand hit at 100-bp resolution; intervals shorter than 1 kb or supported only by hits shorter than 500 bp or below 80% identity were discarded. A query was scored as SSC orientation discordance when the reverse-only evidence was confined to the annotated SSC (at least 80% of reverse-only length), its total length was no more than 1.5 times the SSC, and the largest segment covered at least 40% of the SSC. This term describes the observed orientation state relative to the comparator; it does not assert that a historical isomerization event occurred. When a GenBank record lacked inverted-repeat annotations, the SSC was approximated from the two duplicated rRNA operons; these calls are labelled "rRNA-inferred". Reverse-only evidence spread across the genome was called a complex rearrangement; evidence confined to the inverted repeats was called IR-associated. Comparisons with a best hit below 90% identity, or with the query reverse-complemented over most of its length, were treated as not interpretable. For clade-level suspects, all members of the highlighted clade were tested against a normal accession outside the clade. Every BLAST call was then checked manually against LASTAL dot plots of the same query–comparator pairs (Kiełbasa et al., 2011); where the manual dot-plot reading disagreed with the automatic call, the manual reading was used and is recorded in the per-case table. Per-variant vector dot plots for the manually verified cases are provided in Supplementary Material 2. Scripts and per-query coordinates are provided in `analysis/blast_ssc_ir/` and `scripts/72_blast_ssc_ir.py`.

Manual review and archiving

For each case we rendered a paper-versus-C/A/B panel, highlighted the published long-branch taxon or clade in red, and inspected whether the branch was reproduced by C and removed or shortened by A/B. Cases were classified as strict positive, CDS/PCG-based exclusion, whole-genome non-MSA (local clock or genuine rate heterogeneity), or negative/unconfirmed. Thirty-nine cases were archived for review; `2284410` was later reclassified as negative and excluded from all tree-based statistics. Failed or ambiguous runs were preserved rather than deleted.

Supplementary-material audit

For the 36 strict positives we retrieved the Europe PMC `hasSuppl` flag and the JATS supplementary-material blocks, deduplicated file names, and classified file types. A file was considered machine-readable if it had an alignment, partition or tree extension (FASTA, NEXUS, Newick, PhyloXML, PHYLIP, TXT, XML, ZIP or GZ). The audit captures files registered in the article's supplementary package; author-initiated Figshare deposits outside that package would not be captured.

Statistics

Proportions are reported with Wilson 95% confidence intervals. We used exact one-sided binomial tests against 5% and 1% baselines, Fisher exact tests for year contrasts, and chi-square tests for multi-year heterogeneity. Sensitivity analyses are reported for the two ambiguous classifications (the non-reproduced branch in `2742618` and the two local-clock cases `2622811`/`2629683`).

RESULTS

Corpus and candidate yield

The corpus contained 468 records and 355 tree-bearing papers, of which 335 were branch-length-auditable (Fig. 1). The screen flagged 53 candidates, of which 39 passed review and were archived. The remaining 14 candidates were excluded because they were CDS/PCG-based (8) or because the long branch was attributable to local-clock/real-rate heterogeneity (6).

Final disposition and positivity

Thirty-six papers were strict positives: the published long branch was reproduced by C and disappeared or shortened under A/B (Fig. 2). The confirmed positivity rate was 36/468 = 7.69% (Wilson 95% CI 5.61–10.47%), or 36/335 = 10.75% (7.86–14.52%) among branch-length-auditable papers; counting the 20 matched cladogram-only papers as non-positives gives 36/355 = 10.14% (7.42–13.72%). The positive predictive value was 36/38 = 94.7% (82.7–98.5%) among analysed archived cases and 36/53 = 67.9% (54.5–78.9%) among all candidates. Three archived cases were negative or unconfirmed: `2282225` was a local-clock case, the published branch in `2742618` was not reproduced even by C, and `2284410` was reclassified after re-review because its published branch is shortened rather than eliminated and the original figure already showed the shortened version. The last case is retained in the disposition table but excluded from all tree-based statistics. Among the 36 strict positives, 34 (94.4%) had used direct whole-plastome MSA in the original materials and methods, one used locus-by-locus alignment and one was mixed/self-contradictory. Of the 162 auditable tree-bearing papers that used whole-plastome MSA, 34 were strict positives (21.0%; 95% CI 15.4–27.9%), compared with 2/173 (1.2%; 0.3–4.1%) among papers using locus-by-locus or other strategies (Fisher exact OR 22.71, $P = 8.3 \times 10^{-10}$). This comparison is candidate-enriched rather than an unbiased prevalence estimate.

Ten papers originally classified as "other" were reclassified as whole-plastome MSA on the basis of explicit statements in their figure legends or methods. Five were clear from figure legends or captions: `2466604` ("based on complete plastome sequences"), `2567461` and `2621456` ("based on complete chloroplast genome sequences"), `2301012` ("genome sequences … corrected using the maximum-likelihood (ML) method and the Tamura-Nei model using MEGA 11"), and `2544681` ("ML tree generated in MEGA X using the complete chloroplast genomes"). Five rest on explicit methods statements: `2288889` and `2294896` ("aligned using mafft v7.475"), `2621463` ("based on the complete plastome sequences of 30 Asteraceae species"), and `2694148` ("two datasets—the complete chloroplast genome sequences (Figure 4) and the 62 single-copy gene CDS sequences"), which we count as MSA because the main text reports the whole-genome tree. The tenth, `2420842`, says only "based on the chloroplast gene sequences of *P. deliciosa* and 12 other species"; because the matrix is not specified, we count it as whole-genome but mark it with an asterisk and report a sensitivity analysis without it. One positive, `2438290`, is internally contradictory: the Materials and Methods describe a CDS-based MEGA analysis with K2P, the Results describe 21 whole plastomes aligned with MAFFT, and the figure caption describes 21 complete plastomes. We retain it as a positive but classify its cause separately as "self-contradiction."

Table 1. Final disposition of the 53 candidate papers. The three negatives include `2284410`, which was reclassified after re-review and is excluded from tree-based statistics.

Disposition	*n*
Strict long-branch positive	36
CDS/PCG-based exclusion	8
Whole-genome non-MSA (local clock / real rate)	6
Negative or unconfirmed	3
Total candidates	53

What C, A and B show

In the strict positives, C reproduced the published long branch, whereas A and B removed or substantially shortened it (Fig. 3). Across the 38 analysed archived cases, the median normalized RF was 0.13 (IQR 0.07–0.38) for C–A, 0.17 (0.07–0.38) for C–B and 0.06 (0–0.15) for A–B, indicating that A and B were topologically closer to each other than either was to C. Tree length per site was also method-dependent: the regression of A on C had a slope of 0.44 ($R^2 = 0.39$) and that of B on C a slope of 0.78 ($R^2 = 0.38$); the median log2 ratio was -0.53 for A/C and $+0.90$ for B/C (Fig. 4). Tree length per site is not directly comparable across the three matrices. A is computed on concatenated LCBs, which by construction exclude rearranged and poorly alignable regions and retain only blocks that are collinear across taxa; its alignment is therefore enriched for conserved sequence, and a lower TL/site is expected even if the underlying substitution process is identical. C includes length-variable noncoding regions and regions of uncertain homology, where gaps and mismatches can inflate branch lengths; B is CDS-only, and coding sequence evolves under different constraints, so its higher TL/site is also expected. We therefore treat

global TL as descriptive rather than as evidence that one tree is better. The taxon-level distance residuals control for this composition effect by normalizing each distance matrix and comparing suspect taxa with other taxa within the same matrix; their signal is therefore not explained by the global TL difference. Distance-matrix residuals localized the effect to the suspect taxa: the median residual of curated long-branch taxa was lower than that of other taxa in 26/27 cases for C–A and 25/27 for C–B, but showed no directional difference for A–B (11/27; sign-test $P = 0.44$; Fig. 5). Excluding the non-reproduced case `2742618`, the contrast was 26/26 for C–A and 25/26 for C–B (sign-test $P = 3.0 \times 10^{-8}$ and 8.1×10^{-7}), whereas A–B remained non-directional (10/26; $P = 0.33$). Several cases showed a C-specific clade that was absent from both A and B; others showed near-identical A/B topologies, indicating that LCB and ortholog approaches were converging on the same signal. Root errors and IR/SSC orientation were independent causes: correcting the outgroup or flipping the SSC changed the topology even before alignment strategy was considered. A minority of branches persisted across all three methods; these were retained as non-MSA cases rather than forced into the artifact category.

Supplementary material practices

All 36 strict positives had at least one supplementary file (36/36, 100%; 95% CI 90.4–100%), but none deposited a machine-readable alignment, partition scheme or tree topology (0/36, 0%; 95% CI 0–9.6%). The 58 unique supplementary files were limited to DOCX (31), JPG (14), PDF (7), DOC (4), XLSX (1) and XLS (1). Three papers included spreadsheets, but these were tables rather than alignments; five included PDFs, but these were figures or manuscript files. Some "supplementary" files were clean manuscript copies or language-editing certificates. The bottleneck is therefore not the availability of a supplement or the willingness to use one, but the absence of a requirement for the files that make the analysis reproducible.

Cohort overview

Figure 6 summarizes the flow from year and original alignment strategy through partitioning, multicopy reporting and MEGA use to the final disposition. Before the BLAST re-screen, the attribution panel had a large unexplained component: of the 36 strict positives, 1 was attributable to SSC orientation discordance, 2 to IR expansion, 1 to self-contradictory methods, and 32 remained unattributed. Nine of these 32 had SSC/IR assay data that did not support either mechanism, and 23 had not been assayed. The BLAST re-screen described below resolved most of the latter group.

BLAST re-screen of unassayed cases

Of the 24 unassayed positives, the BLAST screen plus manual LASTAL dot-plot review resolved all 24. Twelve showed SSC inversion: `2290844`, `2294896`, `2306203`, `2309262`, `2347508`, `2383680`, `2412239`, `2437099`, `2466604`, `2505249`, `2528566` and `2635835`. Eleven showed complex rearrangements rather than a simple SSC flip: `2324928`, `2378966`, `2420842`, `2420847`, `2432370`, `2530712`, `2544681`, `2559713`, `2694148`, `2694188` and `2306207`. These include LSC-scale inversions, IR-internal local inversions, whole-sequence reversals and, in several cases, an SSC inversion superimposed on a larger rearrangement (Fig. 7). `2438290`, the self-contradictory paper, also showed an SSC flip, but we retain its separate cause category.

Nine positives had original SSC/IR assays that did not support either mechanism. The LASTAL re-check showed that four were SSC orientation discordance cases (`2281029`, `2288889`, `2567461`, `2621456`) and five were complex rearrangements (`2275827`, `2621463`, `2622811`, `2629683`, `2704257`). The original assays therefore missed LSC inversions, start-point changes and IR-internal inversions.

After adding these manually checked causes, all 36 positives have a structural attribution: 17 SSC orientation discordance cases (1 from the original paper's assay and 16 from the re-screen), 2 IR expansions, 16 complex rearrangements and 1 self-contradictory paper. SSC orientation is the single most common correlate (17/36), followed closely by complex rearrangement (16/36); IR expansion is uncommon (2/36).

We also coded whether the original M&M reported the MAFFT interface. Among the 34 whole-plastome positives, 24 named MAFFT but did not state whether the command line or web server was used; two cited only the MAFFT software URL; two ran MAFFT inside another platform (Geneious Prime or TBtools-II); one reported the command-line `--auto` option; five did not use MAFFT; and none explicitly stated that the MAFFT web server was used. Interface and command-line reporting is therefore almost absent, which is itself a reproducibility gap and prevents attributing the observed orientation patterns to a specific MAFFT front end.

Year patterns

Positivity was 7/35 = 20.0% in 2023, 13/189 = 6.88% in 2024, 8/124 = 6.45% in 2025 and 8/120 = 6.67% in 2026. The 2023 rate was significantly higher than the 2024–2026 rate (7/35 vs 29/433; Fisher exact two-sided $P = 0.012$), whereas the three later years were homogeneous ($\chi^2 = 0.022$, df = 2, $P = 0.989$). Across all four years the heterogeneity was marginal ($\chi^2 = 8.09$, df = 3, $P = 0.044$), and the Cochran–Armitage trend test was not significant ($P = 0.123$). Among branch-length-auditable trees the corresponding rates were 7/29 = 24.1%, 13/138 = 9.42%, 8/86 = 9.30% and 8/82 = 9.76%; 2023 vs 2024–2026 was 7/29 vs 29/277 (Fisher $P = 0.025$), the later years were homogeneous ($\chi^2 = 0.011$, df = 2, $P = 0.995$), and the all-year and trend tests were not significant ($P = 0.114$ and $P = 0.206$). The 2023 sample is small ($n = 35$), so this pattern should be interpreted as hypothesis-generating rather than as evidence of a monotonic decline.

Robustness

If the non-reproduced branch in `2742618` is counted as positive, the rate becomes 37/468 = 7.91% (5.79–10.71%). If the two local-clock cases `2622811` and `2629683` are counted as non-MSA rather than positive, the rate becomes 34/468 = 7.26% (5.25–9.98%). If `2420842` is not counted as whole-genome, the whole-genome association remains (28/169 vs 8/186; Fisher OR 4.42, $P = 1.5 \times 10^{-4}$). Excluding the B-fallback cases does not change the direction of the result.

DISCUSSION

A differential diagnosis of long branches

A long branch is a symptom, not a diagnosis. The same visual pattern can arise from three distinct processes, and the C/A/B design is intended to separate them.

Alignment-driven artifacts

Direct whole-plastome MSA aligns across rearrangements and IR/SSC flips, and gap placement can create long branches. In these cases C reproduces the published branch, whereas A and B remove or shorten it. The pattern is consistent with the standard protocol logic: extract genes or loci, or use LCBs/orthologs, rather than aligning the entire plastome as one block (Bi et al., 2018, 2024; Meckoni et al., 2023; Roestel et al., 2024; Winn et al., 2025; Zhang et al., 2020). With a confirmed rate of 7.7–10.1%, this is not a rare edge case, and the distance-matrix residuals show that the inflation is concentrated in the suspect taxa rather than spread across the tree.

Topology and branch length are not equivalent evidence. The topology returned by ML or Bayesian inference is the tree that maximizes or integrates the joint likelihood over all sites and branches; an incorrect alignment can therefore leave the preferred topology unchanged while distorting the branch lengths that contribute to that likelihood. Simulations of the complete alignment-to-tree process show that MSA error biases estimates of sequence divergence—both individual branch lengths and total tree length—even when the substitution model is correct and the true tree is known, and that these errors are non-randomly distributed across the tree (Md Mukarram Hossain et al., 2015). Alignment errors likewise inflate the false-positive rate of branch-site tests, which depend on branch-specific rate estimates, under a fixed known topology (Fletcher and Yang, 2010; Fletcher, 2011). Model violation adds a second axis: 23.5% of empirical partitions reject stationarity, reversibility or homogeneity (Naser-Khdour et al., 2019), and when violation is imbalanced across the tree, branch-length accuracy can be severely degraded (Naser-Khdour et al., 2026). A stable topology therefore demonstrates that the method tolerates the error; it does not demonstrate that the error is absent. This is why our primary diagnostic is the branch-length residual rather than RF alone: a published tree can pass the topological comparison while retaining exactly the inflated branches that prompted the audit.

Topology is not immune, but the route to topological error is different. A single misaligned sequence usually distorts only its own terminal branch. When several close relatives share the same alignment error—because they share the same SSC/LSC orientation discordance or the same local rearrangement—the erroneous columns provide a shared derived signal whose conditional likelihood can swamp the conditional likelihood signal contributed by the real variation at other sites; the affected taxa are then grouped together (Bergsten, 2005; Md Mukarram Hossain et al., 2015). This is the pattern in 11 positives in which the C-specific clade contained multiple related taxa sharing the same structural variant: `2306203`, `2306207`, `2309262`, `2347508`, `2378966`, `2383680`, `2466604`, `2548850`, `2567461`, `2621463`

and 2629683. In these cases A/B did not merely shorten a terminal branch; they collapsed or redistributed the clade. The practical consequence is that a topology that survives one method is not automatically correct: shared alignment error can produce a supported clade, and only a structural-aware reanalysis or inspection of the underlying alignment will expose it.

Biological rate heterogeneity

Not every long branch is an artifact. Local-clock variation and genuinely fast-evolving lineages exist, and six whole-genome cases in our corpus fell into this category. In these cases the branch persisted across C, A and B and was concentrated in a biologically plausible fast-evolving lineage. The appropriate response is to report the branch and its taxonomic distribution, not to choose a method that erases it.

Rooting and structural errors

Wrong outgroups, IR/SSC orientation and annotation errors can produce the same visual symptom as an alignment artifact. These errors can mislead C, A and B simultaneously, so they must be checked before alignment strategy is evaluated. Quadripartite structure should be standardized and SSC orientation verified (Zhang et al., 2025); the tree should be re-rooted on the paper's stated outgroup for a fair comparison, and any root correction should be reported.

The BLAST re-screen identifies SSC orientation as the dominant structural correlate in this corpus. We use the term SSC orientation discordance rather than SSC isomerization because the comparison establishes a state relative to the comparator, not that a historical isomerization event occurred. Adding the re-screen results, 17 of the 36 positives involve an SSC orientation difference (one confirmed by the original paper's assay and 16 by BLAST/LASTAL), compared with two IR expansions and 16 complex rearrangements. This is mechanistically plausible. A single SSC flip reverses a block of roughly 15–25 kb; a whole-plastome MSA must either align that block in reverse or place gaps around it, and either choice distorts the corresponding branch. IR expansion, by contrast, is rarer and is often visible as a genome-size change before alignment. The practical implication is that checking SSC orientation is not a cosmetic preprocessing step; it is one of the highest-yield diagnostic checks available before a plastome alignment is built.

Diagnostic order. (1) Is the branch reproduced by C? If not, inspect the data, tree file and rooting before invoking MSA. (2) If C reproduces it and A/B remove or shorten it, the balance of evidence favors an alignment artifact. (3) If all three retain it and the branch sits in a fast-evolving lineage, consider real rate heterogeneity. (4) In every case, check outgroup, IR/SSC orientation and annotation first.

Practical recommendations

In practice, we recommend the following procedures: (i) extract PCGs, loci or LCBs rather than aligning whole plastomes as a single block; (ii) align each locus separately, trim the alignments, and concatenate them with an explicit partition scheme; (iii) inspect the alignments for conserved motifs, indels, reading frames and IR/SSC orientation; (iv) use substitution models and partitioning appropriate to the data and report the model-selection procedure; (v) run at least one structural-aware alternative (LCB or orthologs) and compare RF distances and branch lengths; (vi) verify the outgroup and rooting, and test alternative roots; and (vii) report all methods and negative results rather than selecting the tree that "looks best."

Implications for journals and peer review

Fast-turnover taxonomic journals should require a minimal checklist: locus extraction, alignment strategy, partition model, outgroup and IR/SSC handling. Reviewers should ask for alignment files and per-locus trees, not only the final tree. The *Mitochondrial DNA Part B* model—short papers with limited methods review—is particularly vulnerable to these errors.

The infrastructure for a stronger requirement already exists. All 36 positives had supplementary material, but none deposited a machine-readable alignment, partition scheme or tree. *Mitochondrial DNA Part B* is published by Taylor & Francis, which deposits supplementary material on Figshare; Figshare accepts arbitrary file formats and assigns DOIs. There is therefore no technical barrier to requiring, in addition to supplementary figures and tables, the machine-readable alignment matrix, the partition scheme (NEXUS or plain text), and the inferred tree topology with branch lengths and support values (NEXUS, Newick or PhyloXML). These files are by-products of the analysis, cost little to deposit, and would make published conclusions reproducible and falsifiable. Reviewers should be able to inspect these files during review, not only after acceptance.

Limitations

The screen is candidate-based: true prevalence may be higher than the confirmed lower bound because screen-negative papers were not all reanalyzed, and screen sensitivity is unknown. The corpus is restricted to one journal and may not generalize to other journals or taxonomic groups. The B fallback is a name-based single-copy matrix, not equivalent to OrthoFinder; tblastn CDS predictions are deliberately conservative. Manual review was not fully blinded, and the boundary between local-clock variation and an alignment artifact can be subjective. The T92 implementation is a local patch to RAxML-NG and currently supports a single partition; Method A is unpartitioned, so this does not affect the present analyses. The patched model was validated against PAML baseml T92 on a 16-taxon subset: the estimated transition/transversion ratio (3.669) and GC content (0.416) agreed to three decimal places, and both implementations enforced $\pi_A = \pi_T$ and $\pi_C = \pi_G$. The 2023 sample is small. Finally, the supplementary audit captures files registered in the article's Europe PMC/JATS package; an author-initiated Figshare deposit outside that package would not be captured.

CONCLUSIONS

Direct whole-plastome MSA is a systematic and quantifiable source of long-branch artifacts in this corpus, with a confirmed lower bound of 7.7% overall (10.75% among branch-length-auditable papers). The effect is localized: curated long-branch taxa had shorter relative distances in A/B than other taxa in 26/27 C–A and 25/27 C–B comparisons, whereas A–B showed no directional difference. This is a confirmed lower bound rather than a design-unbiased prevalence estimate, but it establishes a recurrent, directional bias rather than idiosyncratic error. SSC orientation was the single most common structural correlate (17/36 positives), followed closely by complex rearrangement (16/36) and IR expansion (2/36); all 36 positives could be assigned a structural cause. LCB- and ortholog-based reanalyses remove or shorten most artifacts, but a minority of long branches are biological. A three-method framework plus alignment inspection provides a defensible, reproducible standard for plastid phylogenomics, and the existing supplementary-material infrastructure is sufficient to make the underlying matrices and trees available.

DATA AND CODE AVAILABILITY

Corpus metadata, the per-case disposition table, RF distances (`analysis/rf_summary_final.tsv`), tree-length summaries (`analysis/tree_length_summary.tsv`), branch-length diagnostics, the supplementary audit, the BLAST re-screen outputs (`analysis/blast_ssc_ir/`), the MAFFT interface audit (`analysis/mafft_web_server/`) and the cladogram-only case list (`analysis/cladogram_only_cases.tsv`) are available with the manuscript. Analysis scripts include `scripts/analysis_pipeline.py`, `scripts/20_tree_comparison_sheets.py`, `scripts/31_build_final_pdf.py`, `scripts/51_tblastn_annotate_missing.py`, `scripts/52_name_single_copy.py`, `scripts/60_make_manuscript_figures.py`, `scripts/64_prelim_rf_mix.py`, `scripts/65_methodA_gtr_freerate.py`, `scripts/66_methodA_t92_freerate.py`, `scripts/68_make_triangle_figures.py`, `scripts/69_branch_residuals.py`, `scripts/70_distance_residual_figure.py`, `scripts/72_blast_ssc_ir.py`, `scripts/73_blast_reattribution.py`, `scripts/74_blast_example_figure.py` and `scripts/78_raxml_t92_methodA.py`. The patched RAxML-NG binary and source patch are in `tools/RAxMLNG-T92/` and `patches/raxml-ng-t92.patch`. The original GTR analyses are archived under `analysis/gtr_original_20261031/`; the IQ-TREE T92 runs used for FreeRate category selection and cross-checks are retained under `data/<dataset>/analysis_v2/methodA_sibeliaz_t92freerate/`. The 38-case supplementary PDF is `analysis/pdf_supp/supplementary_reanalysis_FINAL.pdf`. Supplementary Material 2 (structural-variant dot plots) is `analysis/pdf_supp/supplementary2_structural_variants.pdf`. Supplementary Material 3 (`analysis/sup3/Sup3_reproducibility.tar.gz`) is archived at Zenodo (https://doi.org/10.5281/zenodo.23859886) and contains the analysis scripts, patched RAxML-NG binary and source patches, per-case whole-genome alignments, LCBs, orthoPCG supermatrices and partition schemes, tree files and result tables.

AUTHOR CONTRIBUTIONS

Zhi-Jie Xu conceived the study, performed the literature audit, reanalyses and figure preparation, and wrote the manuscript.

USAGE OF AI-ASSISTANT TOOLS

Claude Code (Anthropic) was used as an AI assistant for batch literature retrieval, monitoring of the batch analysis pipeline, and generation of verification materials (tree panels and diagnostic plots). Every literature record, extracted method statement, and reanalysis output was visually checked by the author against the original article; no result was accepted on the basis of AI output alone. The author is solely responsible for the scientific interpretation and conclusions.

CONFLICT OF INTEREST

The author declares no conflict of interest.

Table 2. Minimum reporting standard proposed for plastid phylogenomic papers.

Item	Requirement
Locus set	PCGs, CDSs, loci or LCBs used; include inclusion/exclusion rules
Alignment	Per-locus alignment files (FASTA or NEXUS); alignment strategy and trimming
Partitioning	Partition scheme (NEXUS or plain text); model selection method
Tree	Tree topology with branch lengths and support (NEXUS, Newick or PhyloXML)
Rooting	Stated outgroup and alternative root tests
Structure	IR/SSC orientation and quadripartite standardization checks
Software	Versions and commands for assembly, alignment and tree inference

Table 3. Supplementary-material audit of the 36 strict positives.

Item	Value
Papers with at least one supplementary file	36/36 (100%; 95% CI 90.4–100%)
Papers with a machine-readable alignment/partition/tree	0/36 (0%; 95% CI 0–9.6%)
Unique supplementary files	58
DOCX	31
JPG	14
PDF	7
DOC	4
XLSX	1
XLS	1

Table 4. Confirmed strict positives by year, using all records and branch-length-auditable trees.

Year	Records	Positives	Rate	Auditable	Rate (auditable)
2023	35	7	20.00%	29	24.14%
2024	189	13	6.88%	138	9.42%
2025	124	8	6.45%	86	9.30%
2026	120	8	6.67%	82	9.76%
Total	468	36	7.69%	335	10.75%

FIGURES

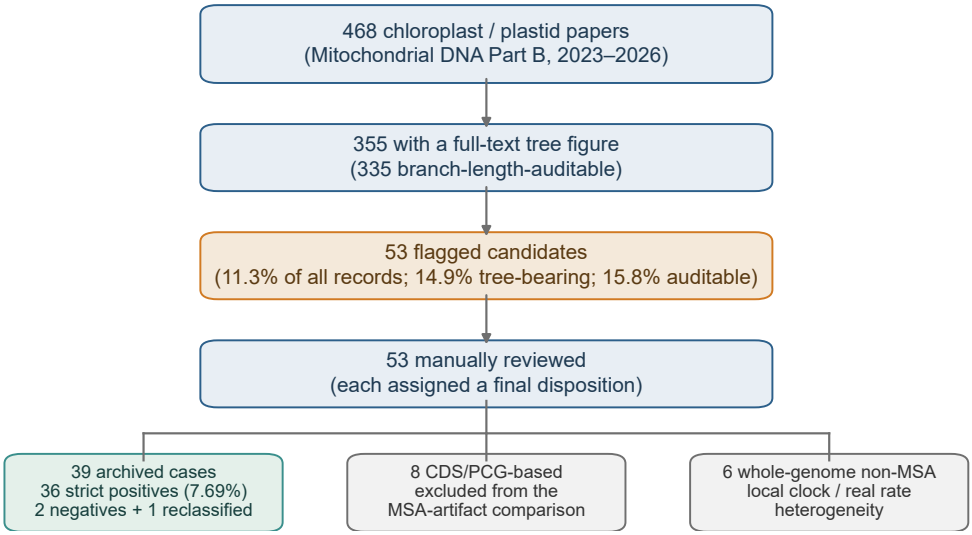


Figure 1. Study profile and final disposition. Of 355 tree figures, 335 were branch-length-auditable and 21 were cladogram-only. The screen is candidate-based: 39 cases were archived, of which 36 were strict positives and 3 were negative or unconfirmed (including one reclassified case excluded from tree-based statistics). The remaining 14 candidates were CDS/PCG-based (8) or whole-genome non-MSA (6).

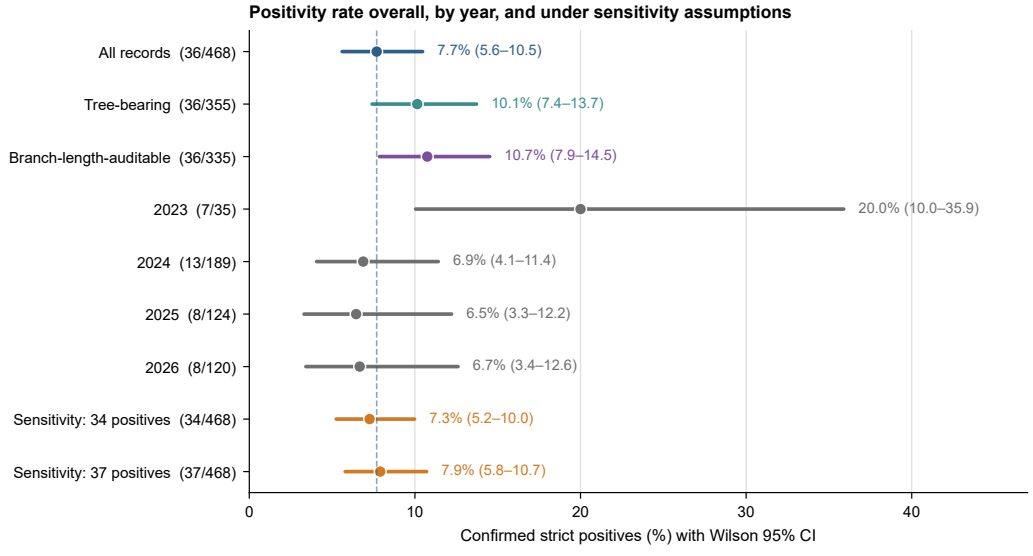


Figure 2. Confirmed positivity rates with Wilson 95% confidence intervals. Blue, all records; teal, tree-bearing records; purple, branch-length-auditable records; gray, year-specific estimates among all records; orange, sensitivity analyses. The dashed line marks the overall estimate (7.69%).

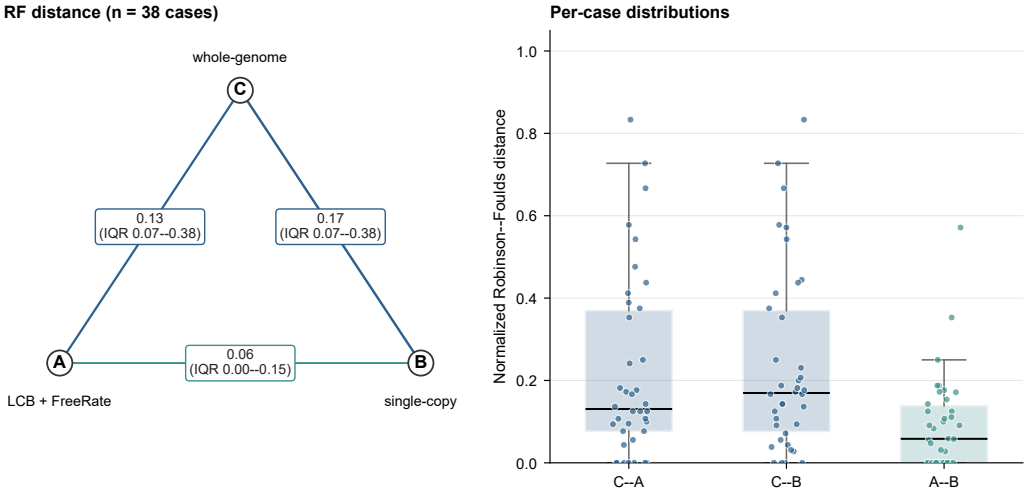


Figure 3. Normalized Robinson–Foulds distances among the three reanalysis strategies for the 38 analysed archived cases. Left: aggregate triangle; edge labels show the median normalized RF and interquartile range. Right: per-case distributions; each point is one dataset.

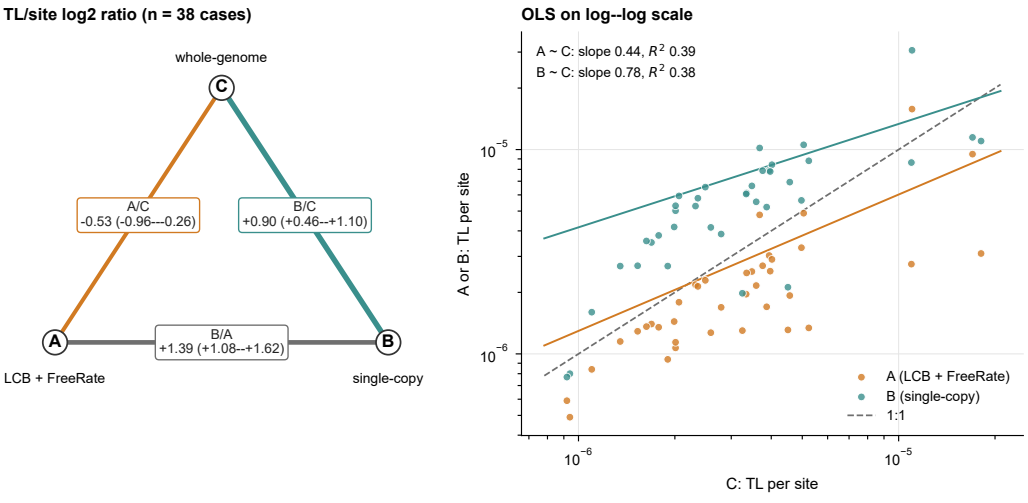


Figure 4. Tree length per site for the 38 analysed archived cases. Left: aggregate triangle; edge labels show the median log2 ratio and interquartile range (positive values mean the numerator has a longer TL per site than the denominator). Right: OLS regressions of A and B on C on log–log axes; the dashed line is 1:1.

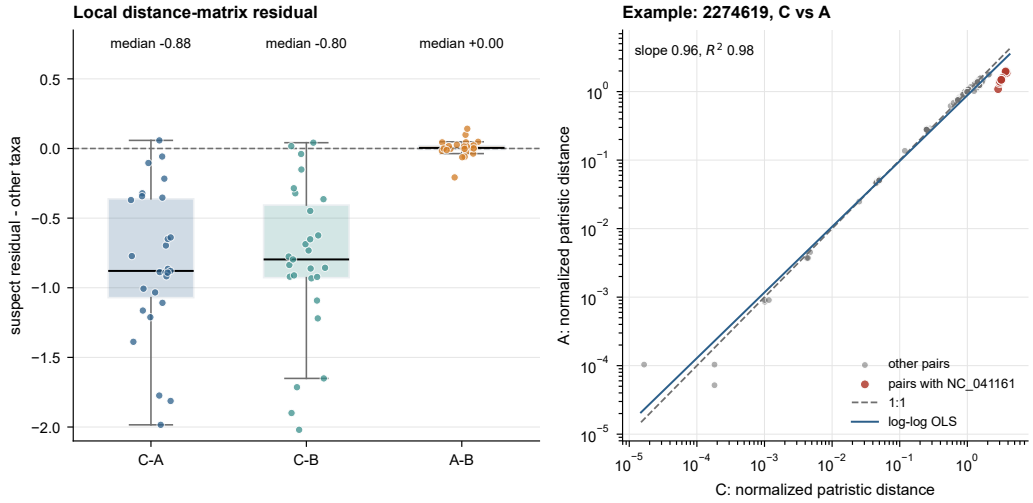


Figure 5. Distance-matrix residuals. (A) Per-case difference between the median residual of curated long-branch taxa and all other taxa for C–A, C–B and A–B; negative values indicate that the suspect taxon is relatively shorter in the second method. (B) Example C–A normalized patristic distance scatter for `2274619`; pairs involving `NC_041161` are highlighted.

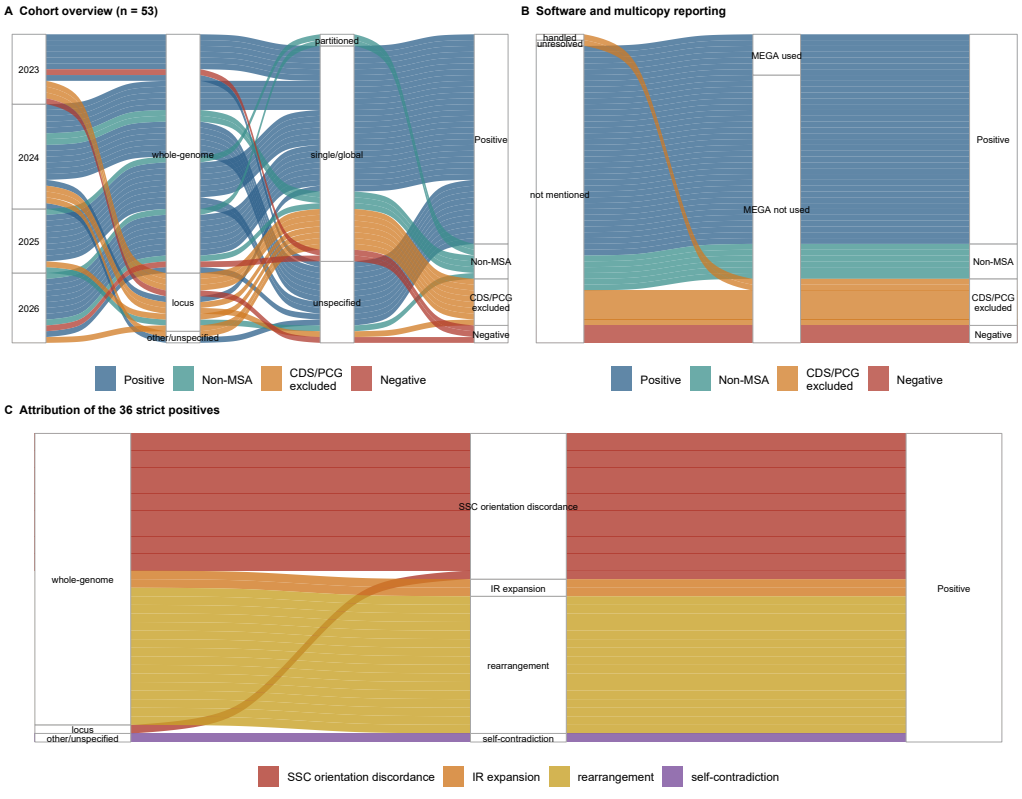


Figure 6. Cohort overview. (A) Year, original alignment strategy, partitioning and final disposition for all 53 candidates. (B) Multicopy reporting and MEGA use. (C) Attribution of the 36 strict positives to SSC orientation discordance, IR expansion, complex rearrangement or self-contradictory methods.

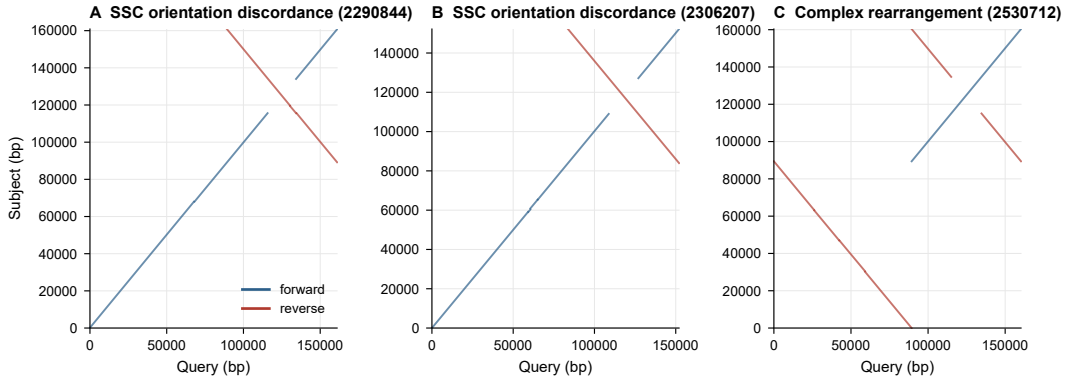


Figure 7. BLAST dot plots from the re-screen of unassayed positives. Blue, forward-strand high-scoring segment pairs; red, reverse-strand pairs. (A) `2290844`: the SSC is reverse-oriented relative to the normal comparator, while the LSC and IRs align forward. (B) `2306207`, representative of the Aster clade: the same SSC-specific reverse block. (C) `2530712`: a complex rearrangement with multiple large reverse blocks, not a simple SSC flip.

REFERENCES

Bergsten, J. (2005). A review of long-branch attraction. *Cladistics*, *21*(2), 163–193. https://doi.org/10.1111/j.1096-0031.2005.00059.x

Bi, G., Mao, Y., Xing, Q., & Cao, M. (2018). HomBlocks: A multiple-alignment construction pipeline for organelle phylogenomics based on locally collinear block searching. *Genomics*, *110*(1), 18–22. https://doi.org/10.1016/j.ygeno.2017.08.001

Bi, G., Luan, X., & Yan, J. (2024). ORPA: A fast and efficient phylogenetic analysis method for constructing genome-wide alignments of organelle genomes. *Journal of Genetics and Genomics*, *51*(3), 352–358. https://doi.org/10.1016/j.jgg.2023.08.006

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421

Capella-Gutierrez, S., Silla-Martinez, J. M., & Gabaldon, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Dong, W., Li, E., Liu, Y., Xu, C., Wang, Y., Liu, K., et al. (2022). Phylogenomic approaches untangle early divergences and complex diversifications of the olive plant family. *BMC Biology*, *20*, 92. https://doi.org/10.1186/s12915-022-01297-0

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*, 238. https://doi.org/10.1186/s13059-019-1832-y

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, *27*(4), 401–410. https://doi.org/10.2307/2412923

Ferrarini, M., Moretto, M., Ward, J. A., Šurbanovski, N., Stevanović, V., Giongo, L., Viola, R., & Cavalieri, D. (2013). An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. *BMC Genomics*, *14*, 670. https://doi.org/10.1186/1471-2164-14-670

Fletcher, W., & Yang, Z. (2010). The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Molecular Biology and Evolution*, *27*(10), 2257–2267. https://doi.org/10.1093/molbev/msq115

Fletcher, W. (2011). *Computational statistics in molecular phylogenetics* [Doctoral thesis, University College London]. UCL Discovery. https://discovery.ucl.ac.uk/id/eprint/1306705

Gitzendanner, M. A., Soltis, P. S., Yi, T. S., Li, D. Z., & Soltis, D. E. (2018). Plastid phylogenomic analysis of green plants: A billion years of evolutionary history. *American Journal of Botany*, *105*(3), 291–301. https://doi.org/10.1002/ajb2.1048

Jin, J. J., Yu, W. B., Yang, J. B., Song, Y., dePamphilis, C. W., Yi, T. S., & Li, D. Z. (2020). GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology*, *21*, 241. https://doi.org/10.1186/s13059-020-02154-5

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). Mod-

elFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., & Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Research*, *21*(3), 487–493. https://doi.org/10.1101/gr.113985.110

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, *35*(21), 4453–4455. https://doi.org/10.1093/bioinformatics/btz305

Kumar, S., Stecher, G., Suleski, M., Sanderford, M., Sharma, S., & Tamura, K. (2024). MEGA12: Molecular Evolutionary Genetic Analysis version 12 for adaptive and green computing. *Molecular Biology and Evolution*, *41*(12), msae263. https://doi.org/10.1093/molbev/msae263

Md Mukarram Hossain, A. S., Blackburne, B. P., Shah, A., & Whelan, S. (2015). Evidence of statistical inconsistency of phylogenetic methods in the presence of multiple sequence alignment uncertainty. *Genome Biology and Evolution*, *7*(8), 2102–2116. https://doi.org/10.1093/gbe/evv127

Meckoni, S. N., Nass, B., & Pucker, B. (2023). Phylogenetic placement of *Ceratophyllum submersum* based on a complete plastome sequence derived from nanopore long read sequencing data. *BMC Research Notes*, *16*, 187. https://doi.org/10.1186/s13104-023-06459-z

Minkin, I., & Medvedev, P. (2020). Scalable multiple whole-genome alignment and locally collinear block construction with SibeliaZ. *Nature Communications*, *11*, 6327. https://doi.org/10.1038/s41467-020-19777-8

Naser-Khdour, S., Minh, B. Q., Zhang, W., Stone, E. A., & Lanfear, R. (2019). The prevalence and impact of model violations in phylogenetic analysis. *Genome Biology and Evolution*, *11*(12), 3341–3352. https://doi.org/10.1093/gbe/evz193

Naser-Khdour, S., Minh, B. Q., & Lanfear, R. (2026). Phylogenetic accuracy under non-stationary and non-homogeneous conditions: A simulation study. *Systematic Biology*, *75*(6), 1196–1206. https://doi.org/10.1093/sysbio/syag010

Qu, X. J., Moore, M. J., Li, D. Z., & Yi, T. S. (2019). PGA: A software package for rapid, accurate, and flexible batch annotation of plastomes. *Plant Methods*, *15*, 50. https://doi.org/10.1186/s13007-019-0435-7

Rambaut, A., Lam, T. T., Max Carvalho, L., & Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution*, *2*(1), vew007. https://doi.org/10.1093/ve/vew007

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, *53*(1–2), 131–147. https://doi.org/10.1016/0025-5564(81)90043-2

Roestel, J. A., Wiersema, J. H., Jansen, R. K., Borsch, T., & Gruenstaeudl, M. (2024). On the importance of sequence alignment inspections in plastid phylogenomics—An example from revisiting the relationships of the water-lilies. *Cladistics*, *40*(5), 469–495. https://doi.org/10.1111/cla.12584

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., & Greiner, S. (2017). GeSeq—Versatile and accurate annotation of organelle genomes. *Nucleic Acids Research*, *45*(W1), W6–W11. https://doi.org/10.1093/nar/gkx391

Timmis, J. N., Ayliffe, M. A., Huang, C. Y., & Martin, W. (2004). Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. *Nature Reviews Genetics*, *5*(2), 123–135. https://doi.org/10.1038/nrg1271

Twyford, A. D., & Ness, R. W. (2017). Strategies for complete plastid genome sequencing. *Molecular Ecology Resources*, *17*(5), 858–868. https://doi.org/10.1111/1755-0998.12626

Wicke, S., Schneeweiss, G. M., dePamphilis, C. W., Müller, K. F., & Quandt, D. (2011). The evolution of the plastid chromosome in land plants: Gene content, gene order, gene function. *Plant Molecular Biology*, *76*(3–5), 273–297. https://doi.org/10.1007/s11103-011-9762-4

Winn, J. C., Bester-van der Merwe, A. E., & Maduna, S. N. (2025). Annotated bioinformatic pipelines for phylogenomic placement of mitochondrial genomes. *Bio-protocol*, *15*(5), e5232. https://doi.org/10.21769/BioProtoc.5232

Wong, T. K. F., Ly-Trong, N., Ren, H., Demotte, P., Banos, H., Roger, A. J., Susko, E., Bielow, C., De Maio, N., Goldman, N., et al. (2026). IQ-TREE 3: Phylogenomic inference

software using complex evolutionary models. *Molecular Biology and Evolution*, *43*(5), msag117. https://doi.org/10.1093/molbev/msag117

Zhang, D., Gao, F., Jakovlic, I., Zou, H., Zhang, J., Li, W. X., & Wang, G. T. (2020). PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Molecular Ecology Resources*, *20*(1), 348–355. https://doi.org/10.1111/1755-0998.13096

Zhang, N. N., Stull, G. W., Zhang, X. J., Fan, S. J., Yi, T. S., & Qu, X. J. (2025). PlastidHub: An integrated analysis platform for plastid phylogenomics and comparative genomics. *Plant Diversity*, *47*(4), 544–560. https://doi.org/10.1016/j.pld.2025.05.005

Supplementary Files

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

- sup2.pdf
- sup1.pdf