# SplitAligner: A Gene-Species Tree Reconciliation Framework Using Split-Based Branch Mapping

Jiaqi Wu[1*]

[1]Graduate School of Integrated Sciences for Life, Hiroshima University, Japan.

[*]Correspondence:

Jiaqi Wu, wujiaqi@hiroshima-u.ac.jp

## Abstract

Phylogenomic analyses are increasingly focused on branch-specific questions within a fixed species tree. However, two pervasive challenges in real datasets—missing taxa and gene-tree/species-tree discordance—complicate the comparability of branches across loci. Here, we introduce SplitAligner, a split-based framework that defines branch identity on a fixed species-tree backbone and evaluates it gene by gene under varying taxon coverage. For each species-tree branch, SplitAligner projects its split onto the gene-specific taxon set to determine whether the branch is evaluable or structurally missing due to a degenerate projected split. Under fixed-topology gene trees, this projection reveals branch fusion, where multiple species-tree branches collapse into an indistinguishable fusion group on the observed taxa. SplitAligner reports such cases as composite fused-branch identities and aggregates branch lengths accordingly. Under free-topology gene trees, SplitAligner further identifies topology-induced missingness (NA_topo), where a branch is decisive under the projected-split criterion but its projected split is absent from the gene tree, separating discordance-driven absence from coverage-driven missingness. These operations produce standardized gene × branch tables and a branch-wise concordance score (Support), defined as the fraction of decisive genes whose free-topology trees recover each projected split. Applying SplitAligner to 2,275 single-copy genes from a dataset of 302 mammals reveals heterogeneous concordance across the mammalian phylogeny and highlights internodes with elevated discordance-associated missingness. The resulting branch coordinate system provides a general framework for branch-based estimates of evolutionary rates, selective constraints, and other branch-wise summaries across thousands of loci/genes.

## Definitions

**Species tree ($S$):** a fixed reference phylogeny representing species relationships, used as the target structure for split mapping.

**Gene tree ($G_g$):** a phylogenetic tree inferred from a single gene or locus $g$, potentially with missing taxa and topological discordance relative to $S$.

**Fixed-topology gene tree (species-tree–constrained gene tree):** a gene tree whose topology is constrained to the species tree; only branch lengths are estimated on this fixed backbone.

**Free-topology gene tree (unconstrained gene tree):** a gene tree inferred without topological constraints.

**Gene-specific taxon set ($T_g$):** the set of taxa present in gene $g$.

**Split (bipartition):** a bipartition of taxa induced by an internal branch of a tree.

**Species-tree split ($\sigma(b)$):** the split induced by an internal branch $b$ of the species tree $S$.

**Gene-tree split:** the split induced by an internal branch of a gene tree $G_g$.

**Split set ($\Sigma(G_g)$):** the set of all gene-tree splits (bipartitions) induced by the internal branches of $G_g$.

**Projected split ($\sigma_g(b)$):** the restriction (projection) of a species-tree split $\sigma(b)$ onto the gene-specific taxon set $T_g$, yielding $\sigma_g(b) = (A_b \cap T_g \mid B_b \cap T_g)$.

**Degenerate projected split:** a projected split that collapses to a trivial partition (e.g., one side becomes empty) due to insufficient taxon coverage.

**Taxon missingness:** missing taxa in a gene tree relative to the species tree.

**Gene×branch cell:** one entry in a gene × branch mapping table for gene $g$ and species-tree branch $b$.

**Structural missingness (NA_struct):** the absence of informative signal for a species-tree branch $b$ in gene $g$ because the projected split $\sigma_g(b)$ is degenerate (the branch is not evaluable for that gene).

**Branch fusion (fusion group):** under a given taxon set $T_g$, multiple distinct species-tree branches can yield identical projected splits, i.e., $\sigma_g(b_1) = \sigma_g(b_2)$, making these branches indistinguishable at the member-branch level.

**Composite fused-branch identity ($b_1|b_2|\ldots$):** an explicit branch label representing a fusion group under a given $T_g$, used as an additional composite row/column to store the fused signal when member branches are indistinguishable.

**Fusion-row missingness (NA_fuse):** the missing label assigned to member-branch rows when $b$ belongs to a fusion group under T_g; the corresponding signal may be represented on an additional composite fused-branch row (e.g., $b_1|b_2|\ldots$).

**Decisive (gene for a species-tree branch):** a gene $g$ is decisive for branch $b$ if $\sigma_g(b)$ is non-degenerate and yields a unique species-branch identity under the fixed-topology baseline (i.e., not labeled as NA_struct or NA_fuse).

**Concordant split:** a decisive projected species-tree split $\sigma_g(b)$ that is present in the corresponding gene tree (i.e., $\sigma_g(b) \in \Sigma(G_g)$).

**Discordant split:** a decisive projected species-tree split $\sigma_g(b)$ that is absent from the corresponding gene tree; its absence implies support for alternative splits in that region (often incompatible with $\sigma_g(b)$).

**Concordant (gene) with branch b:** the gene tree contains the projected split of branch $b$ (i.e., $\sigma_g(b)$ is concordant for gene $g$).

**Discordant (gene) with branch b:** the gene tree does not contain the projected split of branch $b$ despite gene $g$ being decisive for $b$ (i.e., $\sigma_g(b)$ is discordant for gene $g$).

**Topology-induced missingness (NA_topo):** for a decisive gene–branch pair, the mapping is missing under free-topology inference because $\sigma_g(b)$ is absent from the free-topology gene tree (discordance-driven absence rather than coverage-driven non-evaluability).

## 1. Introduction

Phylogenomic studies increasingly ask branch-specific questions on a fixed species tree, yet two pervasive properties of real datasets—gene-tree-discordance and missing taxa—complicate the comparability of branches across loci. Discordance is expected under the multispecies coalescent and is especially pronounced across short internodes, where incomplete lineage sorting generates alternative gene-tree resolutions even without error (*1, 2*). Missing data are similarly widespread in large genomic resources and can affect phylogenomic inference and downstream summaries (*3*). Together, discordance and missing taxa make "branch support" ambiguous unless branch identity and the causes of missing mappings are explicitly defined under gene-specific taxon sets.

A common strategy for enabling branch-wise comparisons across genes is to constrain gene trees to a reference species-tree topology and estimate gene-wise branch lengths on a fixed backbone

(*4*). This provides a convenient coordinate system, but it can also obscure discordance-driven signals by construction and may introduce systematic biases when locus histories deviate from the fixed topology (*5*). Conversely, free-topology gene trees preserve locus-specific histories, but mapping their branches onto a shared species-tree coordinate system becomes non-trivial when missing taxa cause projected branches to collapse or become ambiguous. These limitations motivate an explicit branch-identity infrastructure that remains well-defined under both missing taxa and gene-tree/species-tree discordance.

Several methods summarize concordance across genes, including concordance factors in IQ-TREE2 (gCF/sCF) (*6*) and local branch support measures from coalescent-based species-tree inference (e.g., ASTRAL local posterior probabilities) (*7*). The most closely related split-based tool is PhyParts (*8*), which maps gene-tree bipartitions to a reference species tree to summarize concordant and conflicting splits; it also notes that missing taxa can lead to ambiguous multi-mapping of bipartitions, typically treated as ambiguity or down-weighted support. Here we introduce SplitAligner, which reframes multi-mapping as a projection-induced branch fusion phenomenon and provides a branch-coordinate infrastructure: projected splits define branch identity on gene-specific taxon sets, missing outcomes are decomposed into NA_struct / NA_fuse / NA_topo with an explicit accounting identity, and mapping results are exported as standardized gene×branch tables together with an interpretable branch-wise concordance score (Support) for downstream branch-wise analyses.

We make three contributions.

(1) We define a **split-based branch coordinate system** that preserves branch identity under gene-specific taxon sets, including explicit representation of fused branches under missing taxa.

(2) We provide a **missingness decomposition and accounting framework** (NA_struct / NA_fuse / NA_topo) that separates taxon-coverage effects from discordance-driven absence and yields internal consistency checks for branch-wise mapping.

(3) We introduce a branch-wise **concordance annotation (Support)** that summarizes how frequently each species-tree internode is recovered across free-topology gene trees among decisive loci, enabling direct visualization and quantification of discordance hotspots on the species-tree backbone.

## 2. Overview and Algorithm

### 2.1 Overview of the split-based branch coordinate system

SplitAligner defines branch identity on a fixed species-tree backbone using splits (bipartitions) and evaluates that identity gene by gene under varying taxon coverage. Let $S$ denote the species tree with taxon set $T$, and let $G_g$ denote the gene tree for gene $g$ with observed taxon set $T_g \subseteq T$. Each internal branch $b$ of $S$ induces a species-tree split $\sigma(b) = (A_b \mid B_b)$. To compare this branch across genes with missing taxa, SplitAligner constructs the **projected split** on the gene-specific taxon set,

$$\sigma_g(b) = (A_b \cap T_g \mid B_b \cap T_g).$$

A projected split is **degenerate** if it collapses to a trivial partition (e.g., one side becomes empty) due to insufficient taxon coverage. In that case, the branch is not evaluable for gene $g$, producing **structural missingness**. Otherwise, gene $g$ is **decisive** for branch $b$, and SplitAligner determines whether the projected split $\sigma_g(b)$ is present among the bipartitions induced by the gene tree (i.e., $\sigma_g(b) \in \Sigma(G_g)$). If present, the branch is recorded as **concordant** for that gene; if absent, it is recorded as **discordant** (topology-induced absence under free-topology gene trees).

A key consequence of projection under missing taxa is **branch fusion**. For a given gene $g$, multiple distinct species-tree branches can yield identical projected splits, i.e., $\sigma_g(b_1) = \sigma_g(b_2)$. Such branches form a **fusion group**, meaning they are indistinguishable on $T_g$. Under fixed topology, SplitAligner can report these fused branches explicitly (e.g., as Bs1|Bs3|...) and aggregate branch-length information in a projection-aware manner.

Based on these definitions, SplitAligner produces three standardized outputs. First, it generates **fixed-topology** and **free-topology** branch-assignment/branch-length tables in a common species-branch coordinate system, while preserving branch-level missingness states (structural vs topology-induced). Second, it provides an optional **hybrid (mix-fixed) table** in which topology-induced missing entries in the free-topology table are filled by the corresponding fixed-topology values and explicitly flagged, enabling controlled quantification of discordance-driven missingness. Third, it outputs a **branch-wise concordance annotation** on the species tree, summarizing for each internal branch the fraction of decisive genes whose free-topology gene trees recover the projected split. These outputs together establish an infrastructure for

defining, estimating, and comparing branch-specific quantities across loci under pervasive missing taxa and topological discordance.

**2.2 Algorithm** SplitAligner: six-step split-alignment for gene–species branch mapping

**Input:**

- Gene tree $G$ with branch lengths (optional) and leaf set $L_G$
- Species tree $S$ with leaf set $L_S$
- Optional parameters: maximum merge gap / mapping options (as used in implementation)

**Output:**

- A gene-to-species **branch mapping matrix** $M$, where each gene-tree split is mapped to: (i) a unique species-tree branch, or (ii) a fused species-branch set, or labeled as NA with a reason.

---

**Step 1 — Harmonize taxa (missing-taxa trimming).**

1. Let $L = L_G \cap L_S$.
2. Prune $G$ and $S$ to the shared leaf set $L$, producing $G'$ and $S'$.
3. (Optional) If L is below a minimum threshold, mark this gene as insufficient and stop.

**Step 2 — Extract gene-tree splits.**

4. From $G'$ , enumerate internal branches and represent each branch as a **split** $gi$ (a bipartition of L).
5. Record branch lengths for $g_i$ when available.
6. Store gene splits as an ordered list $G_{splits} = \{g_1,\ldots g_m\}$.

**Step 3 — Extract species-tree splits.**

7. From S' , enumerate internal branches and represent each branch as a **split** $s_j$ (a bipartition of $L$).
8. Store species splits as an ordered list $S_{splits} = \{s_1,\ldots s_n\}$.

**Step 4 — Split alignment (direct mapping).**

9. For each gene split $g_i \in G_{splits}$ :

- if there exists a species split $s_j \in S_{splits}$ such that $g_i = s_j$ (up to side ordering), then set

$$\text{map}(g_i) \leftarrow s_j \text{ (directly mapped)}$$

- Else, set map($g_i$) <- NA_candidate (unmapped at this step).

**Step 5 — Fused-branch resolution (when direct mapping fails).**

10. For each unmapped gene split $g_i$ (NA_candidate):

- Identify a set of **species splits** $F(g_i) \subseteq S_{splits}$ whose union corresponds to the same clade boundary implied by $g_i$ on the pruned taxon set $L$.
- If such a set exists and satisfies the implementation's validity rules, define

$$\text{map}(g_i) \leftarrow \text{union}(F(g_i)) \text{ (fused mapping)}$$

- Otherwise, keep *map($g_i$) <- NA* and assign an NA reason (Step 6).

*(Implementation note: fused-branch length can be computed as the sum of constituent species-branch lengths when all are defined; if any constituent is NA, the fused length is NA.)*

**Step 6 — Assign NA categories and export mapping matrix.**

11. For any $g_i$ still unmapped after Step 5, assign NA labels based on diagnostic rules, e.g.:

- **NA_topo :** topological incompatibility (no valid fused representation on *S'* )
- **NA_struct :** required species branches are structurally absent under the current taxon configuration / missingness
- **NA_fuse :** gene split corresponds to a fused branch where constituent parts are individually NA under the chosen bookkeeping scheme

12. Construct the mapping matrix *M* that records, for each $g_i$ :

- mapped species branch ID or fused-branch ID
- branch length(s) (if used)
- NA label (if applicable)

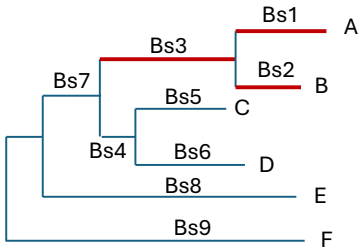
13. Output *M* for downstream counting and decomposition analyses.

# 3. Results

## 3.1 Projection under missing taxa: Fixed-topology mapping under taxon missingness reveals structural missingness and branch fusion

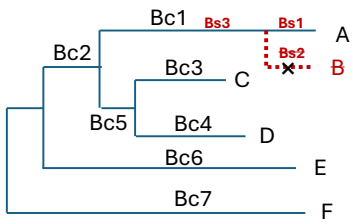
Under pervasive taxon missingness, direct comparison of branches across gene trees is ill-defined because each gene is inferred on a different taxon set. SplitAligner therefore defines branch identity on the species-tree backbone via splits and evaluates each branch $b$ on a gene-specific taxon set $T_g$ using the projected split $\sigma_g(b)$ (Fig. 1A). If $\sigma_g(b)$ is degenerate, the branch is not evaluable for that gene, producing **structural missingness (NA_struct)**. Projection also makes explicit a second consequence of missing taxa: multiple species-tree branches can yield identical projected splits on $T_g$, forming a **fusion group** that is indistinguishable for that gene (Fig. 1B). Under fixed-topology (species-tree–constrained) gene trees, SplitAligner reports the corresponding **fused branch identity** (e.g., Bs1|Bs3) and aggregates branch-length information in a projection-aware manner, while marking the original member branches as **NA_fuse** (Fig. 1B-C). This establishes a consistent branch coordinate system for downstream branch-wise analyses even when terminal taxa are missing.

A. Species tree




***Projection*** : $\sigma_g(b) = (A_b \cap T_g \mid B_b \cap T_g)$
***Gene taxa set***: $T_g = \{A, B, C, D, E, F\}$
***Split***: $\sigma(b) = A_b \mid B_b$
$\sigma(Bs1) = A \mid BCDEF$
$\sigma(Bs2) = B \mid ACDEF$
$\sigma(Bs3) = AB \mid CDEF$

B. Fixed-topology gene tree missing taxa B




***Gene taxa set:*** $T_g = \{A, C, D, E, F\}$
$\sigma_g(Bs1) = \sigma(Bc1) = A \mid CDEF$
$\sigma_g(Bs2) = \emptyset \mid ACDEF$ $*$ *(degenerate)* $\rightarrow$ *NA_struct*
$\sigma_g(Bs3) = \sigma(Bc1) = A \mid CDEF$
***Bc1*** *corresponds to the fusion group* ***{Bs1, Bs3}****, because the projected splits of* ***Bs1*** *and* ***Bs3*** *onto* $\boldsymbol{T_g}$ *are identical and match the gene-tree split at* ***Bc1****.*

***Identity:*** *Bs1|Bs3*
***Length (fixed only):*** *ℓ(Bs1|Bs3)=ℓ(Bs1)+ℓ(Bs3)*

C. Projection under taxon missingness and branch fusion

| Species Tree | Bs1 | Bs2 | Bs3 | Bs1\|Bs3 | Bs4 | Bs5 | Bs6 | Bs7 | Bs8 | Bs9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Fixed-topology output (gene tree) | NA_f | NA_s | NA_f | **Bc1** | Bc5 | Bc3 | Bc4 | Bc2 | Bc6 | Bc7 |

*NA_struct (NA_s): degenerate projected split (structural missingness)*
*NA_fuse (NA_f): branch belongs to a fusion group; fused branch reported as Bs1|Bs3*

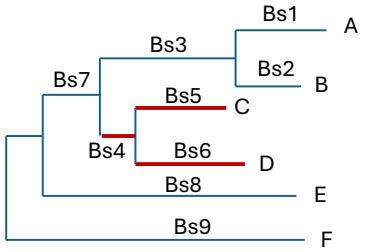
**Figure 1. Projected splits define branch identity under taxon missingness.** Projecting species-tree splits onto the gene-specific taxon set can render some branches structurally missing (degenerate projected splits) and can collapse multiple species-tree branches into an indistinguishable fusion group, represented by a composite branch identity (e.g., **Bs1|Bs3**) under fixed topology. Branches belonging to a fusion group are marked as **NA_fuse** at their original species-branch rows, while the composite identity is reported as an additional branch (e.g., ***Bs1|Bs3***).

### 3.2 Discordance adds topology-induced missing: Free-topology mapping separates topology-induced missingness from structural missingness

Taxon missingness alone does not explain all missing entries in branch-assignment tables when gene trees are inferred with free topology. For a given species-tree branch *b*, if the projected split $\sigma_g(b)$ is non-degenerate, gene *g* is **decisive** for *b* ; in this case, absence of $\sigma_g(b)$ from the free-topology gene tree indicates **topology-induced missingness (NA_topo)** driven by discordance rather than lack of evaluability (Fig. 2A-B). SplitAligner distinguishes this case from **structural missingness (*NA_struct*)**, where $\sigma_g(b)$ is degenerate due to insufficient taxon coverage. This

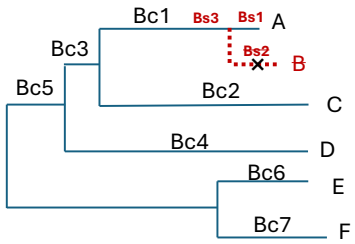
separation is critical because NA_topo is expected to be concentrated on internally discordant regions of the species tree, whereas NA_struct primarily reflects gene-specific taxon coverage (Fig. 2B-C). By retaining these missingness states explicitly in the free-topology table, SplitAligner enables quantitative analyses of discordance-induced bias without conflating it with coverage limitations.

A. Species tree




***Projection*** : $\sigma_g(b) = (A_b \cap T_g \mid B_b \cap T_g)$
***Gene taxa set***: $T_g = \{A, B, C, D, E, F\}$
***Split***: $\sigma(b) = A_b \mid B_b$
$\sigma(Bs4) = CD \mid ABEF$
$\sigma(Bs5) = C \mid ABDEF$
$\sigma(Bs6) = D \mid ABCEF$

B. Free-topology gene tree missing taxa B




***Gene taxa set:*** $T_g = \{A, C, D, E, F\}$
*Decisive:* $\sigma_g\ (b)$ *non-degenerate*
*Discordant (topology-induced missingness):* $\sigma_g(b) \notin \Sigma(G_g)$

$\sigma_g(Bs4) = \cancel{CD|AEF}$ $*$ *(disappeared)* $\rightarrow$ *NA_topo*
$\sigma_g(Bs5) = \sigma(Bc2) = C \mid ADEF$
$\sigma_g(Bs6) = \sigma(Bc4) = D \mid ACEF$

C. Projection under taxon missingness, structural missingness and branch fusion

| Species Tree | Bs1 | Bs2 | Bs3 | Bs1\|Bs3 | Bs4 | Bs5 | Bs6 | Bs7 | Bs8 | Bs9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Free-topology output (gene tree) | NA_f | NA_s | NA_f | Bc1 | NA_t | Bc2 | Bc4 | Bc5 | Bc6 | Bc7 |

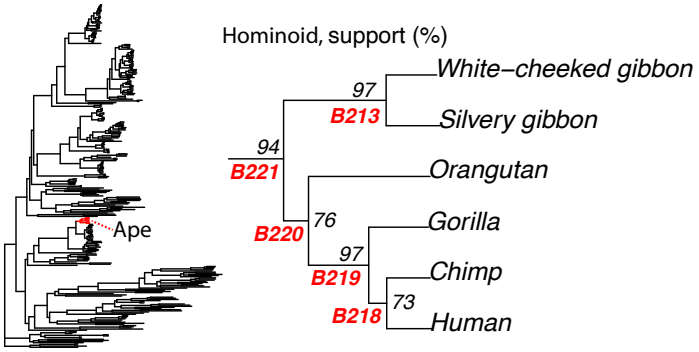
*NA_struct (NA_s): degenerate projected split (structural missingness)*
*NA_fuse (NA_f): branch belongs to a fusion group; fused branch reported as Bs1|Bs3*
*NA_topo (NA_t): topology-induced missingness; the projected species-branch split is absent from the free-topology gene tree despite being decisive.*

**Figure 2. Topology-induced missingness under free-topology gene trees.** Under taxon missingness, projected species-tree splits remain the reference for branch identity; a branch is **decisive** for a gene if its projected split is non-degenerate. Under **free-topology** gene trees, a decisive projected split may nevertheless be absent from the gene tree, yielding **NA_topo** (topology-induced missingness). In contrast, degenerate projected splits produce **NA_struct** (structural missingness), and branches belonging to a fusion group are marked as **NA_fuse** with the fused-branch identity represented by an additional composite branch (e.g., Bs1|Bs3). *Gene trees may also contain additional splits that have no counterpart in the species tree; SplitAligner focuses on branch-wise evaluation of projected species-tree splits in a common backbone coordinate system.*

### 3.3 Branch-wise concordance score and its distribution on internal nodes of a real dataset

To summarize discordance in a branch-resolved manner on the species-tree backbone, we computed a branch-wise concordance score (*Support*) for each internal branch of the 302-mammal species tree. For a given branch $b$ , SplitAligner evaluates, gene by gene, whether the projected species-tree split of $b$ is recovered in the corresponding free-topology gene tree. Because taxon coverage varies across genes, we normalize concordance by the number of genes for which branch $b$ is evaluable under the fixed-topology mapping, yielding $Support(b) = 100 \times n_{unfix}(b) / n_{dec}(b)$. Support is conceptually distinct from bootstrap values or posterior probabilities: it directly quantifies the empirical frequency with which gene trees recover each species-tree branch, conditional on being comparable under missing taxa. We visualized Support on a compact hominoid subtree (Fig. 3A) to provide an externally validated sanity check. At the well-known human–chimpanzee–gorilla divergence, SplitAligner estimates support = 73% for the human+chimpanzee branch, consistent with extensive incomplete lineage sorting at this short internode (Fig. 3B). Extending this annotation genome-wide, Support varies substantially across the mammal phylogeny (Fig. 3C), indicating that discordance-driven absence is not uniformly distributed among internal branches. This branch-wise concordance annotation provides a practical diagnostic for identifying internally discordant regions of the species tree and motivates downstream analyses of topology-induced missingness and bias.

**A. Branch-wise concordance (Support) annotation on an ape subtree**
(by 2275 single-copy genes of 302 mammals)




**B. Concordance calculation table of Ape**

| Branch | Bs1 | Bs2 | Bs3 | Bs4 | Bs5 |
|---|---|---|---|---|---|
| $n_{unfix}(b)$ | 2176 | 2142 | 1724 | 2181 | 1635 |
| $n_{dec}(b)$ | 2240 | 2273 | 2267 | 2252 | 2253 |
| ***Support(b)*** | 97% | 94% | 76% | 97% | 73% |

*Support(b)* = $100 \times n_{unfix}(b) / n_{dec}(b)$

**C. Concordance support distribution of all 299 internal nodes**

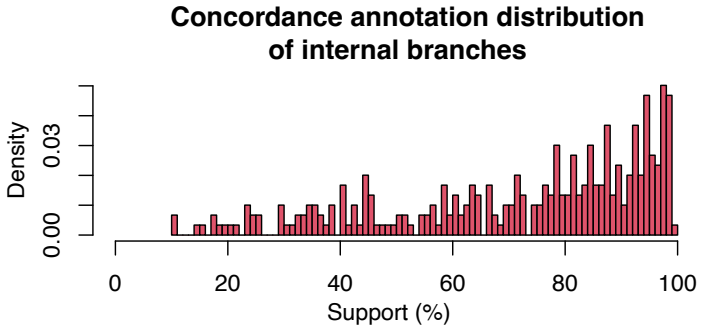
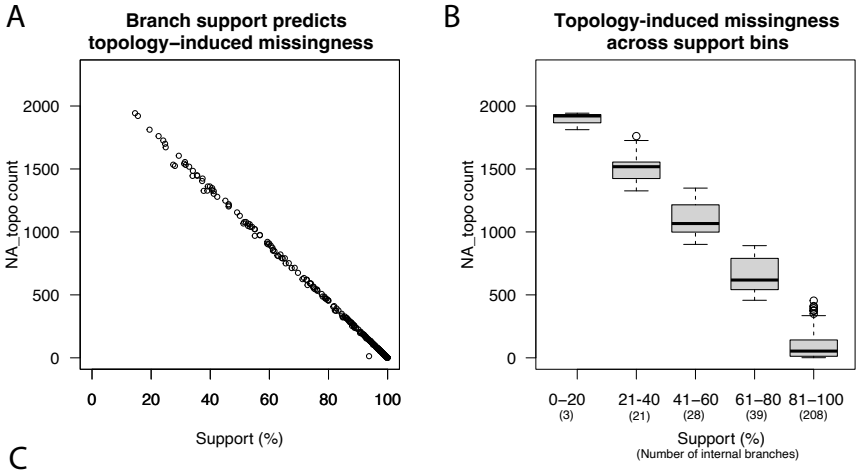



**Figure 3. Branch-wise concordance (Support) annotation on the species-tree backbone.** (A) SplitAligner annotates each internal species-tree branch with a branch-wise concordance score, Support (%), computed from 2,275 single-copy genes in the 302-mammal dataset. A compact hominoid subtree highlights the classic human–chimpanzee–gorilla internode, where the human+chimpanzee branch shows Support = 73%, consistent with extensive incomplete lineage sorting at this divergence. (B) For a given species-tree branch *b*, Support is defined as the fraction of **decisive** genes whose free-topology gene trees contain the **projected split** of *b*; a gene is decisive for *b* if the projected split $\sigma_g(b)$ is non-degenerate on the gene-specific taxon set. (C) Genome-wide distribution of Support across all internal branches of the mammal species tree. Support is conceptually distinct from bootstrap or posterior probabilities because it summarizes empirical gene-tree concordance with each species-tree branch under explicit control of missing taxa.

Building on the branch-wise concordance annotation (Fig. 3), we quantified how gene-tree/species-tree discordance translates into systematic missingness under free-topology mapping. For each species-tree branch b, we distinguished structural missingness (degenerate projected splits) from topology-induced missingness, where a branch is decisive but its projected split is absent from the free-topology gene tree. Across internal branches, topology-induced missingness increases strongly with branch discordance (Fig. 4A), demonstrating that discordance-driven absence is concentrated on specific internodes rather than uniformly distributed. Stratifying branches by concordance further reveals a marked separation in topology-induced missingness (Fig. 4B), supporting the view that free-topology mapping yields branch-dependent evaluability and can induce systematic bias in downstream branch-wise summaries if discordance is ignored. The most discordant internodes (Fig. 4C) provide concrete targets for diagnosing and interpreting discordance hotspots in large phylogenomic datasets.




C

| Branch | Support(%) (n_unfix / n_dec) | m_fix | m_unfix | NA_topo | Annotation * |
|---|---|---|---|---|---|
| B598 | 14.6% (331/2274) | 1 | 1944 | 1943 | MRCA: Artiodactyla, Perissodactyla, Carnivora |
| B597 | 15.5% (353/2274) | 1 | 1922 | 1921 | MRCA: Artiodactyla, Carnivora |
| B541 | 19.5% (438/2250) | 25 | 1837 | 1812 | MRCA: Megaptera, Balaenoptera_musculus |
| B309 | 22.5% (512/2273) | 2 | 1763 | 1761 | MRCA: Lynx, Felis |
| B584 | 24.1% (548/2274) | 1 | 1727 | 1726 | MRCA: Sousa, Globicephala |
| B267 | 24.7% (558/2258) | 17 | 1717 | 1700 | MRCA: Pteropus, Macroglossus |
| B141 | 24.9% (555/2227) | 48 | 1720 | 1672 | MRCA: Glires, Scandentia |
| B475 | 27.5% (581/2115) | 160 | 1694 | 1534 | MRCA: Cephalophus, Eudorcas |
| B476 | 28.0% (594/2118) | 157 | 1681 | 1524 | MRCA: Oreotragus, Cephalophus |
| B509 | 29.4% (667/2272) | 3 | 1608 | 1605 | MRCA: Kobus, Ovis |
| B13 | 31.2% (698/2240) | 35 | 1577 | 1542 | MRCA: Afroinsectivora |
| B178 | 31.7% (711/2243) | 32 | 1564 | 1532 | MRCA: Aotus, Callitrichidae (Callithrix, Saguinus) |
| B585 | 31.4% (713/2268) | 7 | 1562 | 1555 | MRCA: Cephalorhynchus, Globicephala |
| B542 | 32.9% (743/2261) | 14 | 1532 | 1518 | MRCA: Megaptera, Eschrichtius |
| B226 | 34.1% (769/2254) | 21 | 1506 | 1485 | MRCA: Dermoptera, Primates |
| B264 | 34.0% (745/2190) | 85 | 1530 | 1445 | MRCA: Rousettus, Eidolon |
| B137 | 35.5% (798/2248) | 27 | 1477 | 1450 | MRCA: Pedetes, Mus |
| B345 | 35.6% (800/2246) | 29 | 1475 | 1446 | MRCA: Helarctos, Ursus thibetanus |
| B289 | 37.3% (834/2238) | 37 | 1441 | 1404 | MRCA: Rhinoceros, Diceros |
| B85 | 37.4% (851/2275) | 0 | 1424 | 1424 | MRCA: Sciuridae, Hystricomorpha |
| B554 | 37.8% (805/2131) | 144 | 1470 | 1326 | MRCA: Mesoplodon bidens, Mesoplodon europaeus |
| B198 | 39.0% (849/2177) | 98 | 1426 | 1328 | MRCA: Allenopithecus, Erythrocebus |
| B308 | 39.2% (876/2237) | 38 | 1399 | 1361 | MRCA: Felis, Prionailurus |
| B256 | 39.9% (902/2263) | 12 | 1373 | 1361 | MRCA: Artibeus, Tonatia |

*MRCA annotations are given as representative extant taxa bracketing the clade (genus-level unless species-level is required for unambiguous identification).

**Figure 4. Branch support predicts topology-induced missingness across internal branches.** (A) Each point represents one internal branch of the 302-species mammal species tree. The x-axis shows Support (%), defined as $100 \times n_{unfix} / n_{dec}$, where *n_dec* is the number of decisive genes under the fixed-topology baseline and *n_unfix* is the number of those genes that recover the projected split under free-topology mapping. The y-axis shows *NA_topo* count, defined as the excess missing mappings under free-topology relative to fixed-topology ($m_{unfix} - m_{fix}$). (B) Boxplots of *NA_topo* across support bins (0–20, 21–40, 41–60, 61–80, 81–100%), with the number of internal branches in each bin shown in parentheses. (C) Branches with the largest *NA_topo*. For each branch we report Support (%; $n_{unfix} / n_{dec}$), $m_{fix}$, $m_{unfix}$, NA_topo ($m_{unfix} - m_{fix}$), and a representative clade label (MRCA of listed taxa).

### 3.4 Decomposing missingness into structural, fusion, and topology-induced components

Building on the branch-wise concordance patterns (Figs. 3–4), we next decomposed missing mapping outcomes into three mechanistic components: **structural missingness** (NA_struct; degenerate projected splits driven by taxon coverage), **fusion-row missingness** (NA_fuse; indistinguishable adjacent species-tree branches under a gene-specific taxon set, with signal reassigned to composite fused branches), and **topology-induced missingness** (NA_topo; decisive projected splits that are absent from free-topology gene trees). This decomposition provides a concrete accounting of why a species-tree branch can be "missing" under gene-tree discordance and missing taxa, and clarifies how these missingness modes differ between fixed-topology and free-topology mapping.

Across **internal branches**, NA_topo emerged as the dominant difference between fixed and free mapping: NA_topo was absent by definition under the fixed-topology baseline but became abundant under free-topology mapping, whereas NA_struct remained invariant and NA_fuse changed only modestly (Fig. 5A). In contrast, **terminal branches** showed no topology-induced excess missingness, with missingness almost entirely attributable to NA_struct and NA_fuse (Fig. 5B). Consistent with these patterns, internal and terminal counts sum exactly to the genome-wide totals (Fig. 5C), confirming that discordance redistributes mapping outcomes rather than introducing ambiguous bookkeeping. Finally, focusing on all **low-support internodes** (Support < 40%; n=24) highlights that their missingness is overwhelmingly composed of *NA_topo* under free-topology mapping (Fig. 5D), providing a direct mechanistic explanation for the strong support–missingness relationship observed in Fig. 4.

Importantly, the three missingness types form an explicit bookkeeping system. For each mapping cell (gene $g$ × species-tree branch $b$) under a fixed taxon set, projected-split evaluation yields one of four mutually exclusive outcomes: **Mapped**, **NA_struct**, **NA_fuse**, or **NA_topo**. Thus, at the matrix level we have the accounting identity

$$\text{Total cells} = \text{Mapped} + \text{NA_struct} + \text{NA_fuse} + \text{NA_topo}.$$

Moreover, under free-topology mapping, topology-induced missingness reflects a one-to-one replacement of projected species splits by alternative gene-tree splits: for each gene $g$, the

number of **unmapped gene-tree splits** equals the number of **NA_topo** projected splits (i.e., decisive projected splits absent from the gene tree). This provides an internal consistency check that discordance redistributes splits rather than removing signal.

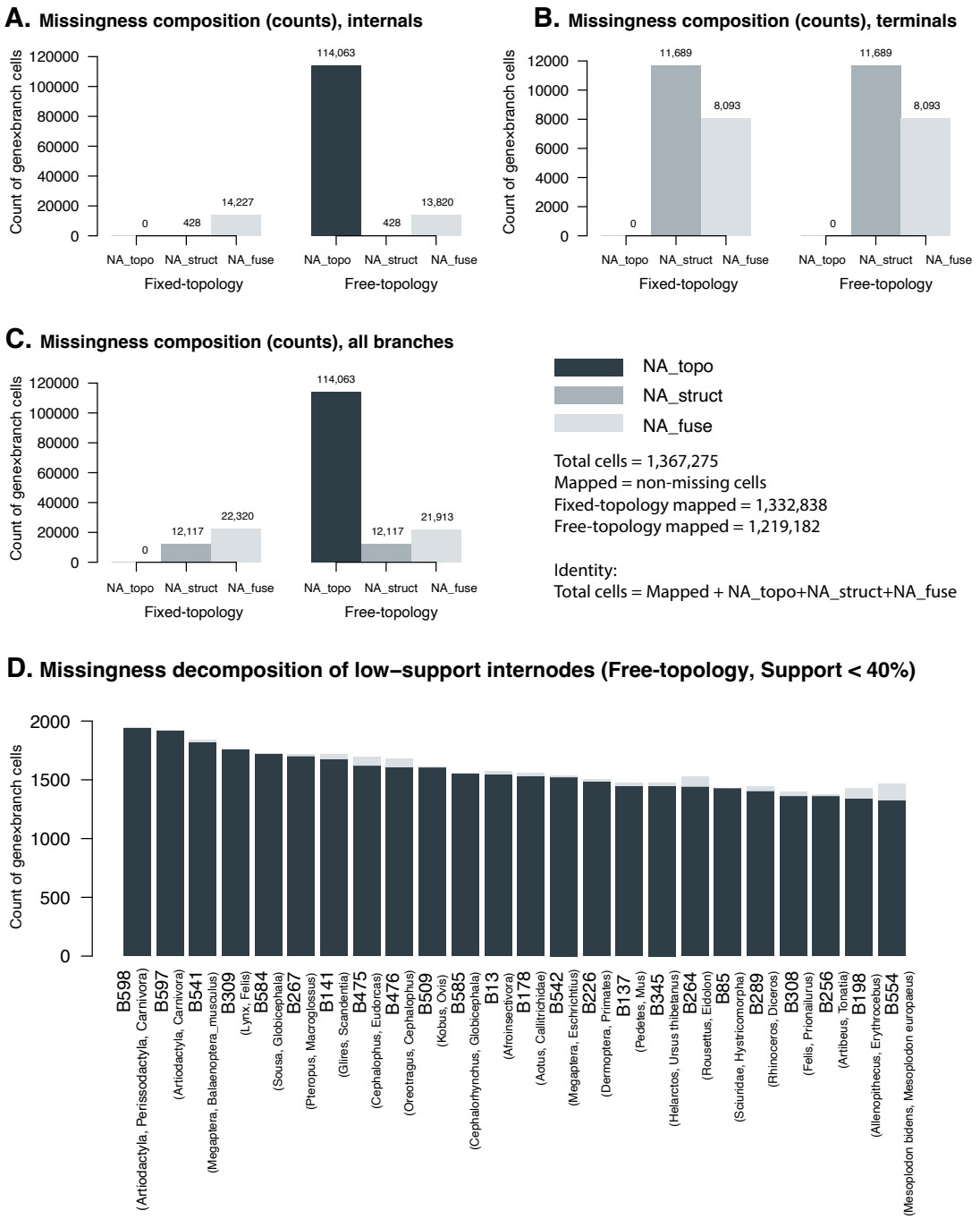



**Figure 5. Decomposing missingness into topology-induced, structural, and fusion components.** (A–C) Missingness composition (counts of gene×branch cells) under fixed-topology and free-topology

mapping, shown separately for internal branches (A), terminal branches (B), and all branches (C). Each cell is classified as Mapped (non-missing), NA_struct (degenerate projected split due to taxon coverage), NA_fuse (fusion-row missingness at original species-branch rows; corresponding signal may be represented on composite fused-branch rows), or NA_topo (decisive projected split absent from the free-topology gene tree). Internal and terminal counts sum exactly to the all-branch totals; the matrix-level accounting identity is shown in-panel. (D) Missingness decomposition for low-support internodes under free-topology mapping (Support < 40%; n = 24). Each bar represents one internode, partitioning gene×branch cell counts into NA_topo, NA_struct, and NA_fuse. Parenthetical labels denote representative clades (MRCA of listed taxa).

# 4. Method

## 4.1 Genome alignments and quality filtering

We downloaded mammalian coding-gene alignments generated by the Hiller Lab (TOGA-based orthology inference on the human hg38 reference assembly) (*9*). The resource contains a catalog of 17,434 coding-gene alignments from 517 placental mammal genome assemblies representing 442 species. To define a consistent species-level taxon set, we retained species with fewer than 1,800 missing genes across the catalog. When multiple assemblies (e.g., subspecies) were available for the same species, we kept a single representative assembly with the fewest missing genes. This procedure resulted in a 302-species placental mammal dataset used in this study.

We further defined a high-confidence single-copy subset for species-tree inference and SplitAligner evaluation, following Wu et al. (2017) (*10*). Specifically, we retained genes with alignment length > 2,000 bp and missing sequences in fewer than 30 of the 302 species, resulting in 2,275 genes, which are released with this study.

## 4.2 Species tree inference

A reference species tree was inferred from the 2,275 single-copy genes. For each gene, we inferred an unconstrained (free-topology) gene tree using IQ-TREE2 with automatic substitution-model selection and partitioning by codon position (1st+2nd versus 3rd positions) (*11*). The species tree was then estimated using ASTRAL under the multispecies coalescent model from the collection of inferred gene trees (*12*).

### 4.3 Fixed-topology branch-length estimation and SplitAligner inputs

SplitAligner takes gene trees and a reference species tree as input and supports both fixed-topology and free-topology mapping. For **fixed-topology** mapping, we constrained each gene tree topology to the reference species tree and estimated **gene-wise branch lengths on the fixed backbone** using BaseML (PAML) (*4*). Unless otherwise stated, branch lengths were estimated under the GTR+Γ+I model (*13-15*). This produces a fixed-topology branch-length table in which branch identity is aligned to the species-tree coordinate system by construction, while allowing gene-specific missing taxa to induce degenerate projections (NA_struct) and fusion groups (NA_fuse).

For **free-topology** mapping, SplitAligner maps unconstrained gene trees (IQ-TREE2) onto the same species-tree branch coordinate system via split-based branch mapping. By comparing fixed-topology and free-topology mapping outcomes, SplitAligner separates missingness due to taxon coverage (NA_struct and NA_fuse) from missingness induced by topological discordance (NA_topo), as quantified in Figs. 4–5.

### 4.4 SplitAligner mapping and output tables

SplitAligner maps gene-tree information onto a fixed species-tree branch coordinate system using split-based branch identities. For a reference species tree and a gene tree with gene-specific taxon set $T_g$, each species-tree internal branch induces a species split, which SplitAligner **projects** onto $T_g$ by restricting both sides of the split to taxa present in the gene. If the projected split becomes trivial due to insufficient taxon coverage (e.g., one side is empty), the branch is not evaluable for that gene and is labeled **NA_struct** (structural missingness).

Under fixed-topology gene trees (species-tree–constrained), projected splits naturally reveal **branch fusion** induced by taxon missingness: multiple adjacent species-tree branches may yield identical projected splits on the same $T_g$. SplitAligner represents such cases by reporting an explicit **composite fused-branch identity** (e.g., B1|B3|…) and marking the corresponding original branch rows as NA_fuse (fusion-row missingness). When branch lengths are used under fixed topology, fused-branch lengths are aggregated from the constituent species-tree branches.

Under free-topology gene trees, a branch can be evaluable (non-degenerate projection) yet its projected species split may be absent from the gene tree due to discordance. SplitAligner labels this outcome as NA_topo (topology-induced missingness), separating discordance-driven absence from NA_struct driven by taxon coverage alone. SplitAligner outputs fixed-topology and free-topology gene×branch mapping tables (TSV), optionally including composite fused-branch columns, together with NA_struct / NA_fuse / NA_topo labels for downstream counting and decomposition analyses.

### 4.5 Branch-wise concordance (Support) and accounting identities

We quantified branch-wise gene-tree concordance on the species-tree backbone using a **Support** score for each internal species-tree branch *b*. Branches labeled as NA_fuse are excluded from **decisiveness** at the member-branch level because missing taxa can collapse multiple adjacent species-tree branches into an indistinguishable fusion group on $T_g$, preventing unique attribution to any single constituent branch. For each gene *g*, SplitAligner evaluates whether branch *b* is decisive under the gene-specific taxon set $T_g$, i.e., whether the projected species-tree split of *b* on $T_g$ is non-degenerate and yields a unique species-branch identity under the fixed-topology baseline (i.e., not labeled as NA_struct or NA_fuse). Among decisive genes, a gene is considered supporting branch *b* if SplitAligner yields a non-missing mapping result on *b* in the free-topology table (equivalently, the projected split of *b* is present in the free-topology gene tree).

Let *N* be the number of genes. For branch *b*, define $m_{fix}(b)$ and $m_{unfix}(b)$ as the numbers of missing mapping outcomes for *b* in the fixed-topology and free-topology tables, respectively. The number of decisive genes for branch *b* is

$$n_{dec}(b) = N - m_{fix}(b) ,$$

and the number of genes that recover the projected split of *b* under free-topology mapping is

$$n_{unfix}(b) = N - m_{unfix}(b).$$

Here, $n_{unfix}(b)$ denotes the number of genes whose free-topology gene trees yield a non-missing mapping on branch *b* (i.e., recover the projected split of *b*). We then define branch-wise concordance (Support) as

$$\mathrm{Support}(b) = 100 \times n_{unfix}(b) / n_{dec}(b),$$

and branch discordance as

$$\text{Discordance}(b) = 100 - \text{Support}(b).$$

In this formulation, Support is a cross-gene concordance measure on a fixed branch coordinate system, conditional on decisiveness under taxon coverage, and is therefore distinct from bootstrap support or ASTRAL-derived branch support measures.

SplitAligner's NA labeling yields an explicit bookkeeping system at the gene×branch matrix level. Each matrix cell belongs to exactly one of four mutually exclusive states: Mapped, NA_struct, NA_fuse, or NA_topo. Accordingly, the total number of gene×branch cells satisfies the accounting identity

$$\text{Total cells} = \text{Mapped} + \text{NA_struct} + \text{NA_fuse} + \text{NA_topo}.$$

In addition, under free-topology mapping, topology-induced missingness corresponds to a one-to-one replacement of projected species splits by alternative gene-tree splits: for each gene under a fixed taxon set, the number of **unmapped gene-tree splits** equals the number of NA_topo projected splits. This relationship provides an internal consistency check that discordance redistributes split support rather than removing signal.

Optionally, we provide a diagnostic hybrid table in which entries missing under free-topology mapping (NA_topo) are imputed using the corresponding fixed-topology estimates and explicitly flagged. This hybrid output is intended for controlled comparisons (e.g., *Fix* vs *Unfix bias quantification*) rather than as a replacement for free-topology mapping.

## 5. Implementation and availability

SplitAligner is implemented as a lightweight, file-based command-line pipeline. The core inputs are a reference species tree (*Newick*) and a set of per-gene gene trees (*Newick*; one tree per locus), and the core outputs are branch-mapping matrices in tab-delimited format for downstream analyses.

**Core workflow.** For each gene, SplitAligner (i) extracts splits from the gene tree and the species tree, (ii) projects species-tree splits onto the gene-specific taxon set, (iii) performs split matching to map gene-tree splits onto the species-tree branch coordinate system, and (iv) reports mapped branches, composite fused branches (when applicable), or missing outcomes. The pipeline

444 produces two primary branch-mapping tables: a **fixed-topology table** (branch lengths estimated
445 on a fixed backbone) and a **free-topology table** (unconstrained gene trees), which together
446 enable separation of taxon-coverage-driven missingness (NA_struct / NA_fuse) from
447 discordance-driven missingness (NA_topo).

448 **Outputs.** SplitAligner writes (1) gene×branch matrices for fixed-topology and free-topology
449 mapping, (2) optional composite fused-branch columns (e.g., B1|B3) for explicit representation
450 of fusion groups, and (3) optional species-tree annotations (e.g., branch-wise Support). Output
451 tables are plain-text TSV files designed for easy counting and visualization.

452 **Optional post-processing utilities (for Fig. 5 decomposition).** We provide two small utilities
453 that operate on the SplitAligner output matrices without altering the underlying split mapping:
454 Extract_NA_fuse.pl labels fusion-row missing entries (NA_fuse) based on the presence of
455 composite fused-branch columns, and Confirm_NA_structure.pl labels structural missingness
456 (*NA_struct*) and topology-induced missingness (NA_topo) by cross-referencing fixed- and free-
457 topology matrices under identical branch headers. These utilities are optional and are used here
458 to generate the missingness decomposition shown in Fig. 5.

459 **Computational complexity.** Given precomputed splits, per-gene mapping scales approximately
460 linearly with the number of genes and the number of splits per gene; fusion resolution is
461 evaluated only when direct matching fails. The implementation is suitable for large
462 phylogenomic datasets with thousands of loci.

463 **Availability.** Source code and documentation are available at
464 https://github.com/wujiaqi06/SplitAligner.

465 ## 6. Discussion

466 **SplitAligner addresses a practical but often implicit problem in phylogenomics: branch**
467 **identity becomes unstable when gene trees and species trees differ in both topology and**
468 **taxon sampling.** By formulating branch mapping in a split-based coordinate system,
469 SplitAligner makes branch comparability explicit under gene-specific taxon sets. In particular,
470 the notion of **projected species-tree splits** provides a deterministic basis for evaluating whether
471 a gene tree can be meaningfully compared to a given species-tree internode, even when many
472 taxa are missing. This resolves a common ambiguity in downstream branch-wise analyses, where

473 “missing” values are frequently treated as a single category despite arising from distinct
474 mechanisms.

475 A central outcome of this framework is the explicit **decomposition of missingness** into
476 NA_struct, NA_fuse, and NA_topo, together with a bookkeeping identity at the gene×branch
477 matrix level. NA_struct captures degeneracy of projected splits driven by taxon coverage,
478 NA_fuse captures cases where multiple adjacent species-tree branches become indistinguishable
479 after projection and are represented by composite fused branches, and NA_topo captures
480 discordance-driven absence of decisive projected splits on free-topology gene trees. This
481 decomposition clarifies why free-topology mapping can introduce systematic, branch-specific
482 missingness beyond taxon coverage alone, and it provides internal consistency checks that help
483 distinguish discordance effects from incomplete sampling or implementation artifacts.

484 SplitAligner’s **Support** annotation complements, rather than replaces, traditional measures of
485 phylogenetic confidence such as bootstrap support or ASTRAL-derived local measures.
486 Bootstrap quantifies stability under resampling within a single-tree model, whereas Support
487 summarizes **cross-gene concordance on a fixed species-tree branch coordinate system**,
488 conditional on decisiveness under taxon coverage. In this sense, Support is directly aligned with
489 the needs of branch-wise comparative analyses across loci: it highlights internodes that are
490 repeatedly recovered as projected splits across free-topology gene trees and anticipates where
491 discordance is expected to manifest as topology-induced missingness (NA_topo).
492 **Recommended usage** is therefore to interpret Support jointly with the missingness
493 decomposition—especially on internal branches—rather than treating all missing mappings as
494 equivalent. SplitAligner is related to split-based summaries such as PhyParts and concordance-
495 factor approaches (e.g., gCF/sCF), but differs in explicitly defining branch identity under gene-
496 specific taxon sets and decomposing missingness into NA_struct/NA_fuse/NA_topo under a
497 matrix-level accounting identity (*6-8*). In particular, SplitAligner interprets missing-taxa–driven
498 ambiguous multi-mapping as branch fusion with explicit composite identities, enabling
499 standardized gene×branch tables for downstream analyses rather than node-centric concordance
500 summaries.

501 Notably, the lowest-support internodes identified by SplitAligner correspond to several well-
502 known “difficult” nodes in mammalian phylogeny that have been repeatedly discussed in the

503 literature, including deep placental splits and rapid radiations within Laurasiatheria and
504 Cetartiodactyla, as well as lineage-rich clades such as primates, bats, and odontocetes. These
505 internodes are not merely low-confidence in a generic sense: SplitAligner shows that their
506 apparent instability is accompanied by elevated NA_topo under free-topology mapping,
507 indicating that discordance manifests as topology-driven absence of projected species splits
508 rather than as simple taxon-coverage loss (NA_struct) or fusion-row bookkeeping (NA_fuse). In
509 this way, SplitAligner provides a mechanistic, branch-resolved decomposition of where and how
510 discordance erodes branch comparability across loci, offering a practical diagnostic layer
511 complementary to bootstrap or coalescent-based support values. We emphasize that SplitAligner
512 does not re-infer phylogenetic relationships for these nodes; rather, it provides an infrastructure
513 for making branch identity and branch-wise discordance/missingness explicit under
514 heterogeneous taxon sampling.

515 Finally, we note practical considerations and extensions:

516 - **Scope and assumptions.** SplitAligner assumes a fixed reference species tree and gene
517 trees inferred on (possibly different) taxon subsets. Mapping is performed at the level of
518 splits and is intended as a branch-coordinate infrastructure rather than a
519 network/reticulation inference method.

520 - **Interpretation of fusion.** Fusion is reported as a branch-row bookkeeping construct:
521 NA_fuse indicates indistinguishability of adjacent species-tree branches under projection,
522 with signal potentially reassigned to composite fused branches. Users should interpret
523 NA_fuse accordingly when summarizing branch-level signals.

524 - **Extensions.** The projected-split coordinate system and accounting framework can be
525 extended to alternative reconciliation rules, additional diagnostics, or explicit network
526 representations, while preserving the key goal of comparable branch identity under
527 missing taxa and discordance.

528 Future work may leverage the distribution of unmapped gene-tree splits (complements of
529 NA_topo under fixed taxon sets) to characterize alternative split support around highly
530 discordant internodes, including classical three-taxon conflicts.

531

532 ## References


533 1. W. P. Maddison, Gene Trees in Species Trees. *Systematic Biology* **46**, 523–536 (1997).
534 2. J. H. Degnan, N. A. Rosenberg, Gene tree discordance, phylogenetic inference and the
535 multispecies coalescent. *Trends Ecol Evol* **24**, 332-340 (2009).
536 3. H. Philippe *et al.*, Resolving difficult phylogenetic questions: why more sequences are
537 not enough. *PLoS Biol* **9**, e1000602 (2011).
538 4. Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**,
539 1586-1591 (2007).
540 5. F. K. Mendes, M. W. Hahn, Gene Tree Discordance Causes Apparent Substitution Rate
541 Variation. *Syst Biol* **65**, 711-721 (2016).
542 6. B. Q. Minh, M. W. Hahn, R. Lanfear, New Methods to Calculate Concordance Factors
543 for Phylogenomic Datasets. *Mol Biol Evol* **37**, 2727-2733 (2020).
544 7. E. Sayyari, S. Mirarab, Fast Coalescent-Based Computation of Local Branch Support
545 from Quartet Frequencies. *Mol Biol Evol* **33**, 1654-1668 (2016).
546 8. S. A. Smith, M. J. Moore, J. W. Brown, Y. Yang, Analysis of phylogenomic datasets
547 reveals conflict, concordance, and gene duplications with examples from animals and
548 plants. *BMC Evol Biol* **15**, 150 (2015).
549 9. B. M. Kirilenko *et al.*, Integrating gene annotation with orthology inference at scale.
550 *Science* **380**, eabn3107 (2023).
551 10. J. Wu, T. Yonezawa, H. Kishino, Rates of Molecular Evolution Suggest Natural History
552 of Life History Traits and a Post-K-Pg Nocturnal Bottleneck of Placentals. *Curr Biol* **27**,
553 3025-3033 e3025 (2017).
554 11. B. Q. Minh *et al.*, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic
555 Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534 (2020).
556 12. C. Zhang, M. Rabiee, E. Sayyari, S. Mirarab, ASTRAL-III: polynomial time species tree
557 reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**, 153 (2018).
558 13. S. Tavaré, Some Probabilistic and Statistical Problems in the Analysis of DNA
559 Sequences. *Lectures on Mathematics in the Life Sciences* **17**, 57–86 (1986).
560 14. Z. Yang, Maximum likelihood phylogenetic estimation from DNA sequences with
561 variable rates over sites: approximate methods. *J Mol Evol* **39**, 306-314 (1994).

15. J. S. Shoemaker, W. M. Fitch, Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. *Mol Biol Evol* **6**, 270-289 (1989).

## Data availability

All data products generated in this study—including the 2,275 single-copy gene trees, gene×branch mapping tables with NA_struct/NA_fuse/NA_topo labels, branch-wise Support summary tables (including the full table underlying Fig. 4C), and the annotated species tree—are available at the SplitAligner GitHub repository (https://github.com/wujiaqi06/SplitAligner) in standard plain-text formats (e.g., *Newick*, *TSV*), together with scripts to reproduce the figures and key analyses.

## Author Contributions

J.W. conceived the study, developed the SplitAligner algorithm, implemented the software, performed all computational analyses, generated the figures, and wrote the manuscript.

## Competing interests:

The author declares no competing interests.

## Acknowledgement:

This study was supported by Grant-in-Aid for Scientific Research (B) 23K23950 from the Japan Society for the Promotion of Science. I thank ChatGPT (OpenAI) for assistance with language editing, manuscript organization, and figure-caption polishing. All scientific decisions, analyses, and interpretations were made by the author.