

Where the tree of life is empirically resolved, and where it is not: an open atlas of species-level phylogenies and their archival uncertainty

Francisco Richter¹ [ORCID](#)

¹ Università della Svizzera italiana (USI), Lugano, Switzerland

✉ richtf@usi.ch

Comparative and macroevolutionary biology depends on knowing two things at once: where the species-level tree of life is empirically resolved, and where it is not. Yet the phylogenies that anchor downstream inference are routinely treated as fixed inputs, even though their empirical substrate is heterogeneous in ways that vanish once a tree is downloaded. Trees differ in inference framework, in dating regime, and in the balance between species supported by direct molecular data and species inserted by taxonomic constraint. They also differ in whether per-node uncertainty estimated during inference survives the path from primary analysis to deposited derivative. These choices accumulate silently into cross-clade analyses, and no existing resource maps the current state of phylogenetic knowledge uniformly across Bacteria, Archaea, and Eukaryota.

We offer a starting point: a first systematic inventory of what is, and is not, empirically resolved. We assemble an atlas of 264 species-level phylogenies, mapped onto a single global species dictionary of roughly 637,000 labels and organised into 62 partitions of described biodiversity, with 36 further partitions explicitly retained as unrepresented so that phylogenetic dark matter remains visible at the same resolution as the represented clades. For every canonical tree, a per-tree provenance ledger records inference framework, dating regime, the fraction of tips anchored in molecules rather than in taxonomy, and an audit of how much inferred uncertainty survived deposition; an explicit sensitivity envelope of inclusion criteria, from permissive to strictest, brackets every coverage statement against its own auditable rule. Because knowledge of the tree of life is itself moving, the atlas is released as a versioned, continuously updated resource through the companion R package *phyloatlas* and a per-partition wiki, so that downstream comparative analyses can propagate provenance and inclusion-criterion uncertainty into their own inferences rather than silently committing to one set of choices.

Introduction

Phylogenetic trees are the organizing principle of comparative biology. From estimating diversification rates¹ to prioritizing species for conservation², from reconstructing ancestral traits³ to understanding biogeographic patterns⁴, nearly every macroevolutionary question depends on having a species-level phylogeny. Over the past two decades the number of molecular phylogenies has grown rapidly, driven by high-throughput sequencing and improved phylogenomic methods⁵.

Yet despite this progress, the body of available phylogenetic data is heterogeneous in ways that constrain its use as a comparative substrate. Trees vary in their inference framework (Sanger supermatrices, ultraconserved-element and anchored-hybrid-enrichment phylogenomics, transcriptome-based multi-species coalescent inference, supertree aggregation) and in their dating regime (Bayesian relaxed-clock methods, penalised likelihood with *treePL*, compiled published chronograms re-dating, fossilised birth-death tip-dating, or no time calibration at all). They also vary in the balance between species with direct molecular sequence support and species inserted through taxonomic constraints or synthesis-pipeline grafting, a difference that is consequential for downstream comparative inference but is rarely visible in the partial, non-standardised metadata that researchers actually pull together when assembling trees from multiple sources for analysis.

The same trees are deposited under inconsistent archival conventions: per-node uncertainty (HPD intervals, posterior samples, bootstrap replicates) is preserved in some derivatives and flattened to a point estimate in others, with little relationship to the original analysis. Species labels are reconciled against different taxonomic backbones (Catalogue of Life, GBIF, NCBI Taxonomy, ITIS, clade-specific authorities), so the same name does not always resolve to the same concept across studies. The practical consequence is that researchers running cross-clade comparative analyses must either trust opaque provenance or assemble the relevant per-tree details by hand from primary sources. Existing infrastructure addresses individual axes of this heterogeneity. The Open Tree of Life⁶ provides a single global synthesis of ~2.4 million tips but at the cost of replacing ~93% of tips with taxonomic descent rather than empirical phylogenetic input. TreeBASE⁷ and TreeHub⁸ archive published trees, through voluntary author deposit or automated literature extraction, but do not curate clade-level representatives or expose comparable per-tree provenance. VertLife⁹ curates phylogenies for tetrapods but does not extend to plants, invertebrates, or microbes. What is missing is a coordinated cross-kingdom view in which every retained tree's inference framework, dating regime, taxonomic-placement fraction, and archival-uncertainty status are documented uniformly and side by side.

Here we present an open, standardised atlas of empirical species-level phylogenies that takes this heterogeneity as its starting point. Each retained tree is curated, mapped to a global species dictionary, and annotated with per-tree provenance: inference framework, dating regime, the fraction of tips with direct molecular sequence support versus tips placed by taxonomic constraint or synthesis-pipeline grafting, and the archival status of per-node uncertainty in the deposited file. The atlas distinguishes three kinds of species at the tip level. *Molecular tips* are species with direct molecular phylogenetic support: a tip backed by sequence data in the source study. *Taxonomic tips* are species present in atlas trees but placed by taxonomic constraint or synthesis-pipeline grafting: a tip inferred from the catalogue rather than the data. *Unrepresented species* are those recognised by authoritative taxonomic catalogues but absent from every atlas tree. Beyond all three sits the broader pool of *estimated true biodiversity* (described and undescribed species combined), which we also track, because the same coverage percentage can mean very different things depending on whether the denominator is the catalogued count or the projected total.

We organise the assessment around 62 atlas partitions of described biodiversity. The partitions are *data-driven* rather than rank-based: each is defined as the largest taxonomic unit for which a single tree could plausibly serve as the partition-canonical resource at the current state of the literature. Where one comprehensive tree covers a class (Mammalia, Insecta), the class is a single partition; where a class has only sub-clade trees (Arachnida is split into spiders, scorpions, mites and ticks, and a residual “other arachnids”), the sub-clades become separate partitions. The scheme is also *dynamic*. Partitions migrate from the unrepresented set to the represented set as new phylogenies are published; they may split when one comprehensive tree gives way to several specialised ones, or merge when a new megatree subsumes prior clade-level efforts. The 62-component snapshot reported here is therefore the empirical picture as of May 2026, not a fixed taxonomic claim, and the atlas is released as a versioned, continuously-updated collection that we maintain on an ongoing basis as new trees become available.

The remainder of the paper reports what this audit reveals: the coverage asymmetry across kingdoms, the time-calibration gap that concentrates in invertebrate and microbial clades, the archival-uncertainty gap in which per-node uncertainty rarely survives the path from inference to deposit, and the three explicit sensitivity bounds that quantify how much of the headline coverage depends on accepting taxonomically placed tips.

Materials and Methods

Definitions

We use the following tip-type terminology throughout. A *molecular tip* is a species present in an atlas tree with direct molecular phylogenetic support, i.e., backed by sequence data in the source study. A *taxonomic tip* is a species present in an atlas tree but placed by taxonomic constraint or synthesis-pipeline grafting rather than by its own sequence

data. An *unrepresented species* is a species recognised by an authoritative taxonomic catalogue but absent from every atlas tree. Beyond all three sits the broader pool of *estimated true biodiversity* (described plus undescribed species), which we also track because the same coverage percentage can mean very different things depending on which denominator is used.

For bookkeeping we additionally retain three operational terms. A *dataset* is a single downloadable phylogenetic resource from one study or database (one row in Table 2). A *taxonomic group* is the biological clade covered by one or more datasets (e.g., “Birds” is one group covered by two datasets). An *atlas partition* is the largest taxonomic unit for which a single tree could plausibly serve as the *partition-canonical tree*: the single tree per partition used for coverage calculations, typically the one with the most species-level tips. The partition scheme is *data-driven* (defined by the empirical granularity at which broadly representative species-level trees do or could exist, not by taxonomic rank or backbone topology) and *dynamic* (partitions migrate, split, and merge as new phylogenies appear; see §Updating cadence and versioning). In the bookkeeping above, “representative tree” is synonymous with “partition-canonical tree”.

Trees in the atlas are classified into three categories with respect to how their topology and tip-set were assembled; this assembly trichotomy is the qualitative counterpart of the per-tree `otl_pipeline_involvement` flag recorded in Supplementary Table S5 (direct empirical phylogenies carry *none*; empirical syntheses carry *full-pipeline*, *collection-backed*, or *taxonomy-backed-grafting* according to how synthesis software or grafting was used). (i) *Direct empirical phylogenies* are inferred from molecular or genomic data in a primary study (e.g.,⁹ for mammals,¹⁰ for sharks). (ii) *Peer-reviewed empirical syntheses* are clade-bounded trees built by combining published input phylogenies, sometimes augmented with taxonomic placement of species lacking sequence data, and released as standalone peer-reviewed deliverables with a primary methodology paper (e.g.,¹¹ for birds,¹² for seed plants,¹³ for insects); these may use software pipelines such as the Open Tree of Life synthesis algorithm as an algorithmic backbone, but the resulting tree is published, citable, and clade-bounded. (iii) The live *Open Tree of Life synthesis product* (~2.4 million tips, of which only ~7% are placed by phylogenetic input trees and the remainder derive from taxonomic descent) and any other tree dominated by automated taxonomic placement without a clade-bounded primary methodology paper are excluded. Categories (i) and (ii) are both included; per-tree provenance, including the molecular-tip fraction and any Open Tree pipeline involvement, is reported in Supplementary Table S5.

For orientation, five counting units appear at different points in this paper, all derived from the same underlying collection as of May 2026: **49 datasets** (Table 2 rows; one source study or database per row) $\geq$ **47 taxonomic groups** (distinct clades; two groups have two datasets each, three rows are cross-kingdom) $\subseteq$ **62 atlas partitions** (the complete partition of described biodiversity used for coverage assessment in Fig. 1; 26 with a shipped canonical tree, 36 unrepresented) and **264**

standardized tree files in the public archive. These 264 files comprise 218 Condamine family-level trees¹⁴, 25 further partition-canonical trees, and 21 sub-clade and reference trees; the 26th shipped canonical, *Crocodylia*, is itself one of the 218 Condamine trees, so $218 + 25 + 21 = 264$ files cover the 26 shipped canonicals without double-counting, all mapped through a single **global dictionary of 637,619 standardized species labels**. Because the atlas is rebuilt periodically (§Updating cadence and versioning), these counts are a snapshot rather than a fixed total.

What counts as empirical synthesis

The distinction between included empirical syntheses (category ii) and the excluded synthesis product (category iii) turns on four operational tests. A category-(ii) tree must be (1) *clade-bounded*: it covers a named taxonomic clade rather than the live tree of life as a whole; (2) accompanied by a *primary methodology paper* describing its construction; (3) a *peer-reviewed deliverable* in its own right; and (4) a *citable downloadable artifact* (a specific, version-pinned tree file in a public repository, not a continuously-updated database snapshot).¹¹ is an empirical synthesis: it uses Open Tree synthesis software (Chronosynth, `python-opentree`, `addTaxa`) but draws approximately 85% of its 10,824 bird tips (i.e., molecular tips) from 281 published input phylogenies across 262 studies, with the remaining ~1,585 species placed via Clements 2021 taxonomy as taxonomic tips; the result is a peer-reviewed bird tree, not a fragment of the live OTL product.¹² is an empirical synthesis: approximately 22% of its 356,305 seed-plant tips are molecular tips with direct GenBank sequence data, with the remainder grafted via Open Tree of Life taxonomy against a molecularly-resolved backbone (ALLMB); the published tree is itself the primary methodology paper, not a database export. The live ~2.4M-tip Open Tree of Life synthesis is excluded under this criterion because the great majority of its tips are placed by taxonomic descent and it is released as a continuously-updated database rather than as a clade-bounded peer-reviewed deliverable. Supplementary Table S5 records per-tree molecular-tip fraction and Open Tree pipeline involvement so the inclusion of each canonical can be audited against this checklist.

Partitioning the tree of life

We partition described biodiversity into 62 components defined by the empirical granularity at which broadly representative species-level trees do or could exist. The partition is *data-driven, not biologically motivated*: each component is the largest taxonomic unit for which a single tree could plausibly serve as the partition-canonical resource. Where one comprehensive tree covers a class (Mammalia, Insecta), the class is a single partition; where a class has only sub-clade trees (Arachnida is split into Spiders, Scorpiones, Acari, and a residual “Other arachnids”), the sub-clades become separate partitions. Residual buckets (“Other arachnids”, “Other hexapods”, “Other Mollusca”, “Other vertebrates”, “Other phyla”) ensure that every described species sits in exactly one partition. The 36 unrepresented partitions (Supplementary Table S3) are those for which no broad downloadable

species-level molecular phylogeny was identified; this makes the phylogenetic dark matter visible at the same resolution as the represented clades.

A taxonomic-rank carving (by class, by order, by family) would not produce comparable coverage measurements across the tree of life, because the same rank corresponds to vastly different amounts of data-availability heterogeneity in different lineages. A phylogenetic carving from a backbone tree would presuppose a topology, which is inappropriate in a paper whose subject is the gaps in those topologies. The data-driven partition makes “% coverage” a well-defined, uniformly comparable quantity across all 62 components and acknowledges that the partition itself is an empirical artifact reflecting current tree availability, not a normative taxonomic claim. The 62-component framing is also *dynamic*: a partition undergoes a *split* when a new clade-bounded comprehensive tree appears for a sub-clade of an existing partition (the sub-clade becomes its own partition, with the parent re-scoped accordingly), and a *merge* when a new cross-clade comprehensive tree subsumes two existing partitions into one. Partitions also migrate from the unrepresented to the represented set as new phylogenies are published, without changing the partition definition. These split/merge/migrate events are detected and applied as part of each rebuild (§Updating cadence and versioning).

Literature survey and curation

Because the partition scheme is data-driven, the literature survey itself shapes partition boundaries: when a search returns a comprehensive new tree for a sub-clade of an existing partition, the partition splits; when a cross-clade megatree appears, two partitions merge. We surveyed the phylogenetic literature through May 2026 to identify the most comprehensive, species-level, empirical phylogenetic trees for each major taxonomic group. For each group, we searched for “[taxon] phylogeny species tree” and related queries across Google Scholar and targeted web searches, complemented by backward citation tracking from recent comparative studies. Repository searches were conducted on Dryad, Zenodo, Figshare, and GitHub. We also searched clade-specific databases (FishTreeOfLife, insectphylo.org, BeeTree, GTDB, VertLife, 10kTrees, TreeBASE) and bioRxiv for preprints deposited in 2025–2026.

Trees were required to be: (1) inferred from molecular or genomic data (not taxonomy alone), (2) species-level (or genome-cluster level for prokaryotes), (3) published or from an established database with published methodology, and (4) freely downloadable from a public repository. When multiple candidate trees existed for a group, we selected the tree with the greatest number of species-level tips; alternative trees were retained as secondary datasets. Trees behind journal paywalls or lacking deposited data files were excluded. Non-English resources were not systematically searched but were included when encountered. This search strategy prioritized the largest available downloadable trees rather than exhaustively cataloguing every published phylogenetic placement in the literature. The search is rerun at each rebuild (§Updating cadence and versioning), so the May 2026 cut-off described

here is a snapshot rather than a final state.

Validation

All tree files were validated using the *ape* package v5.0¹⁵ in R. Validation checks include successful parsing (`read.tree` or `read.nexus`), tip count verification (`Ntip`), branch-length presence, ultrametricity assessment (`is.ultrametric` with tolerance 0.1), and crown age extraction (`max(branching.times(tree))`) for dated trees. We treat *dating* and *ultrametricity* as logically independent properties: a tree is *dated* if its branch lengths are in units of time (typically millions of years), and *ultrametric* if all root-to-tip distances are equal. The two properties usually coincide in the atlas because most clade-level chronograms in the published literature are extant-only and so happen to be ultrametric, but the coincidence is not definitional: fossilised birth-death and total-evidence chronograms with fossil tips are dated but not ultrametric (Upham et al. 2019 publishes both an extant-only NDexp chronogram and a parallel FBD tip-dated chronogram), and a tree with branch lengths in substitutions per site under a strict clock can be ultrametric without being dated. The dated/undated flag in Supplementary Table S5 therefore follows the source publication's documentation of branch-length units: "dated" if the source paper states the deposited tree is time-calibrated, "undated" if branch lengths are in substitutions per site or absent. Ultrametricity is recorded as a separate auxiliary flag in the per-partition metadata. Multifurcations were retained as-is. Rooting was accepted as provided by the original study. Failures were handled deterministically: trees that fail parsing are logged and excluded from the release; trees that the source paper documents as time-calibrated but that fail ultrametricity (e.g., FBD or total-evidence chronograms with fossil tips) are retained as dated with the non-ultrametric flag set, not reclassified as undated; trees whose tip count differs from the source paper's reported count by more than 1% are flagged for manual review. The validation script is invoked once per rebuild (§Updating cadence and versioning).

Standardization

Tip labels were standardized to Genus_species format through the following steps: (1) trailing family/order annotations in ALL CAPS were stripped (e.g., *Zaglossus_bartoni_TACHYGLOSSIDAE* → *Zaglossus_bartoni*); (2) specimen voucher codes and institutional identifiers were removed; (3) subspecies trinomials were reduced to binomials; (4) spaces and hyphens between genus and species were normalized to underscores. GTDB genome accessions were retained as-is because they serve as species-cluster identifiers. Labels that could not be resolved to binomials (e.g., environmental samples) were retained but flagged.

After cleaning, duplicate species within a single tree were collapsed to a single representative standardized label for the archive. The global dictionary was then constructed by sorting all unique standardized labels alphabetically and assigning sequential numeric IDs (1 to 637,619). All trees were rewritten in Newick format with numeric tip labels. To

enable interoperability with other biodiversity databases, all dictionary entries were mapped against the Catalogue of Life 2025 checklist¹⁶: 83.2% of matchable eukaryotic names were resolved to CoL accepted names (352,185 exact matches and 58,016 via synonym resolution), with the mapping deposited alongside the standardized collection.

The CoL mapping is a first-class atlas product. Where a standardised label resolves to multiple CoL accepted names (typically because the source tree uses an older binomial that has since been split), we keep the synonym-resolved primary match as the canonical mapping and record the alternates in an `alternates` field in the exported dictionary, so the ambiguity is visible rather than silently collapsed. The dictionary is regenerated from scratch at every rebuild (§Updating cadence and versioning) and exported as a downloadable CSV alongside the standardised tree files; the CSV is the authoritative source for label-ID mappings between releases.

The global total of 637,619 therefore reflects deduplicated standardized labels across the full collection. By contrast, the group-level coverage percentages reported in Table 2 and the summary figures follow the partition-canonical tree totals used for each source study or database resource, which are typically species-level tips or species-cluster representatives as distributed by the original study.

Per-tree provenance (Supplementary Table S5)

Supplementary Table S5 is the per-tree provenance ledger that anchors the empirical-synthesis inclusion test (§What counts as empirical synthesis), the permissive/molecular/strict coverage bounds (§Coverage under explicit inclusion criteria), the coverage assessment (§Coverage assessment), and the crown-age audit (§Crown-age uncertainty extraction). Each row corresponds to one partition-canonical tree in the atlas.

Each row records the tree's provenance across roughly two dozen columns; four are load-bearing for the analyses in this paper, and they cleanly separate the two independent axes of tree provenance. The first two describe placement. `molecular_tip_fraction` is the proportion of tips in the deposited tree backed by direct molecular sequence data in the source study, with the complementary fraction comprising taxonomic tips. `otl_pipeline_involvement` is a categorical flag (`none`, `full-pipeline`, `collection-backed`, or `taxonomy-backed-grafting`) describing whether and how Open Tree of Life synthesis software or taxonomy-backed grafting contributed to the shipped tree. The remaining two encode the two provenance axes we return to throughout. The *dating method* (how node ages were inferred) is recorded in `dating_code`, a five-value code (BRC Bayesian relaxed clock, PL penalised likelihood, ST supertree, SYN dated synthesis, N undated). Independently, the *archival class* (which representation of crown-age uncertainty was actually deposited) is recorded, for the trees carried into the crown-age audit, in the companion inventory `fig2_hpd_inventory.csv` under the column `hpd_method`, with four shipped values: `beast_mcc_annotation` (a BEAST MCC tree carrying per-node `height_95%_HPD` annotations), `posterior_sample` (a deposited sample of trees from the posterior), `bootstrap_treepl` (a deposited

bootstrap-replicate distribution), and none (a single dated tree with no accompanying uncertainty, or an undated tree). Dating method sets the epistemic ceiling on what uncertainty could exist; archival class records what survived to the deposited file. `imputation_status` flags the source-tree provenance for placement: `none`, `partial`, or `heavy`.

The audit procedure is bounded and reproducible. For each canonical tree we read the source paper, its supplementary methods, the repository README, and any comments embedded in the deposited NEXUS or Newick file; we record the evidence for each of the four fields verbatim alongside the value; cases where the evidence is ambiguous or contradictory are flagged for re-review by a second curator. The audit boundary is explicit: we did not contact authors, did not re-fit chronograms, and did not search the body of supplementary PDFs for HPDs that were absent from the deposited tree files.

Historical succession of canonical trees

The deposited file `table_s7_canonical_succession.csv` (in the Zenodo and GitHub v1.0.8 release) is the temporal counterpart of Supplementary Table S5: where S5 records what the current atlas ships, the succession ledger records the historical succession of canonical trees per partition from 2006 to 2025. Each row documents one canonical-at-time tree with its publication year, citation key, Crossref-resolvable DOI, tip count, molecular-tip fraction, dating code, and Open Tree pipeline involvement, plus a `canonical_in_atlas` flag that is `yes` for the single row whose DOI matches the atlas-shipped canonical for that partition and `no` for predecessors and for post-atlas-canonical literature trees that the atlas has not (yet) adopted. The table contains 95 rows across the 26 represented partitions, with 26 entries flagged `canonical_in_atlas=yes` (one per partition) and 69 documenting prior canonicals or methodologically distinct alternatives. The curation procedure mirrors the S5 audit boundary: every entry was Crossref-verified before acceptance, candidates without a resolvable DOI were rejected, and confidence is recorded per row as `high` (Crossref + full text read), `medium` (Crossref + abstract only), or `low` (Crossref hit but tip count or molecular fraction not extractable from the deposited material). The table supports two downstream uses: (i) reproducing historical states of the atlas at any year between 2006 and 2025, enabling longitudinal analyses of coverage growth and methodological turnover that go beyond the snapshot reported here; and (ii) documenting that the atlas's canonical choices are auditable judgement calls rather than the most-recent-paper-wins default. As an example of the latter, mammals retains Upham et al. 2019 (2019) as the atlas canonical despite the more recent Álvarez-Carretero et al. 2022 (2022), because the Upham tree's full taxonomic coverage (5,912 tips vs ~4,700 tips in the Bayesian timetree) better serves comparative-methods users at the class level; S7 documents the timetree as a non-canonical alternative.

Coverage under explicit inclusion criteria

Several large phylogenies in our collection extend molecular backbones to unsampled species using taxonomic placement

algorithms. In the Smith & Brown¹² seed plant tree, approximately 79,881 of 356,305 tips (~22%) are molecular tips with GenBank sequence data; the remainder are taxonomic tips placed using Open Tree of Life taxonomy. Similar approaches were used for squamates¹⁷ (~4,161 of 9,755 molecular) and sharks¹⁰ (~610 of 1,192 molecular); these full TACT-imputed trees are shipped in the standardized release. The Rabosky et al. 2018 fish tree as published contains 31,516 tips of which only 11,638 (~37%) are molecular tips; for fish we ship the molecular-only subset (`actinopt_12k_raxml`; 11,638 tips) in the standardized release rather than the full TACT-imputed tree, and Table 2 reports the molecular-only subset's coverage (32.2%, 11,638 of ~36,000 described ray-finned fish). For comparability with the published Rabosky et al. tree, the source study's full-tree coverage was 87.5% (31,516 / 36,000 described species), of which 32.2% derives from direct molecular data and the remainder from TACT-imputed placements. We follow⁹ in distinguishing *strict type-1 placements* (a tip backed by its own sequence data) from *type-2 placements* (a tip placed under a genus- or family-level taxonomic constraint without its own sequence data) when reporting per-tree molecular-tip fractions.

We include the three TACT-imputed trees (seed plants, squamates, sharks) because they represent the most comprehensive species-level phylogenies available for each group and are the community-standard reference. The fish case is asymmetric (we ship the molecular-only subset rather than the full tree) because the full fish tree is numerically dominated by taxonomic placements rather than serving as a comprehensive synthesis, whereas the seed-plant and squamate TACT-imputed trees are the canonical published reference for downstream comparative work. Coverage percentages for the three shipped TACT-imputed trees therefore reflect total tips (including taxonomic placement) rather than species with independent molecular phylogenetic support. We therefore report headline coverage under three operationally explicit inclusion criteria: **permissive** (all retained canonical tips), **molecular** (molecular-only on the three TACT-imputed canonicals: seed plants, squamates, sharks), and **strict** (additionally excluding full Open Tree of Life pipeline trees), computed from the per-tree molecular-tip fractions documented in Supplementary Table S5 and summarised in Table 1.

Deduplicated unique-label counts are derived from the summed source-tree tips by collapsing duplicates across partitions and applying name resolution against the Catalogue of Life 2025 checklist (§Standardization). For the permissive-criterion numerator, the 474,342 summed eukaryotic source-tree tips resolve to 494,005 unique standardised labels after this step: the count grows rather than shrinks because the name-resolution step admits accepted-name plus distinct-synonym mappings within the dictionary structure used by the atlas. For the molecular and strict criteria the same accounting is propagated to each criterion's molecular-tip subset, giving the headline figures of ~211,000 (molecular) and ~114,000 (strict) unique standardised labels reported in Table 1. Note that the name-resolution lift differs between criteria because synonym/accepted-name pairs are concentrated in the high-volume taxonomically-grafted regions excluded

Table 1 | Three sensitivity bounds on global eukaryotic coverage. All three criteria use the same denominator: ~2.1 M described eukaryotic species (sum of described-species counts across the 62 atlas partitions, Supplementary Table S5).

Inclusion rule	Numerator	%
All retained canonical trees in full, including taxonomic tips.	494,005 unique eukaryotic standardised labels	23.7%
Molecular tips only on the three TACT-imputed trees (seed plants, squamates, sharks); other trees counted in full.	~211,000 (excludes ~283,000 imputed tips)	~10%
Additionally exclude full OTL-pipeline trees (birds ¹¹ , insects ^{13,19}); for remaining partial-placement trees (mammals ⁹ , amphibians ²⁰ , sharks ¹⁰ , squamates ¹⁷ , plants ¹²), count molecular tips only.	~114,000 unique eukaryotic standardised labels with direct molecular support (Fig. 2)	~5.4%

under stricter criteria; the permissive-criterion ratio is therefore not a uniform multiplier for the molecular and strict criteria. The three bounds together make the empirical placement gap explicit at the resolution of operational choices, and Supplementary Table S5 lets any reader recompute coverage under their own preferred inclusion criterion. Users requiring strictly molecular phylogenies can use the GenBank-only tree variants available from the original studies.

Several partitions are additionally covered by one or more *alternative trees* (recorded in the deposited file `table_s6_alternative_trees.csv`) that are not the partition-canonical choice but are retained because they serve roles the canonical alone cannot. First, sub-clade trees within a partition provide finer-resolution sampling (Primates, Carnivora, Cetaceans within Mammals; the Portik et al. 2023 frog tree within Amphibians; the Kawahara et al. 2023 butterfly chronogram within Insects) and let trait-evolution studies work at the resolution of interest rather than at the class level. Second, methodologically distinct alternatives are available: Jetz et al. 2012 alongside McTavish et al. 2025 for birds; Pyron et al. 2013 alongside Tonini et al. 2016 for squamates; PAFTOL²⁵ alongside Smith & Brown¹² for seed plants. These let comparative-methods users assess topology- and method-sensitivity by re-running their analysis on both trees and reporting the empirical disagreement. Third, the alternates serve as *substitution targets* for users who define stricter inclusion criteria than the strict criterion by substituting a non-OTL-pipeline alternative wherever an OTL-pipeline canonical is excluded; the permissive, molecular, and strict bounds reported here do not perform this substitution by default (under the strict criterion the affected partitions are dropped), but the alternative-trees CSV (`table_s6_alternative_trees.csv`) documents the available substitution candidates so any user-defined inclusion rule can be applied without leaving the framework.

Coverage assessment

For each taxonomic group, coverage was calculated as the partition-canonical tree total reported for that dataset divided by the estimated total described species for that group, using only groups with a directly comparable denominator. Denominators were obtained from authoritative taxonomic databases accessed in May 2026: the Mammal Diversity Database²⁶, IUCN Red List, Reptile Database²⁷, Amphibian Species of the World²⁸, AmphibiaWeb, Eschmeyer's Catalog of Fishes, Catalogue of Life¹⁶, World Flora Online, World Spider Catalog, Species Fungorum, World Register of Marine

Species, GTDB²⁹, and AlgaeBase. Per-database access dates and version identifiers are listed in Supplementary Table S4.

The global coverage figure of ~23.7% reported in the abstract and introduction (the permissive criterion; see §Coverage under explicit inclusion criteria) was calculated as the number of unique eukaryotic standardized labels in the standardized dictionary (494,005) divided by the total described eukaryotic species across all groups surveyed (~2.1 million). The numerator is deduplicated: each standardized label is counted once regardless of how many trees contain it. The denominator sums described-species counts from authoritative sources for each independent top-level group (e.g., Insects rather than its constituent orders), plus described-species estimates for the 36 unrepresented partitions documented in Supplementary Table S3 and the main table footnotes. Some datasets in the collection cover subgroups of larger clades (e.g., Butterflies within Insects, Orchids within Seed plants); these contribute to the atlas but their described-species counts are not added to the global denominator separately from their parent group.

The same numerator can also be reported against *estimated true biodiversity* (described plus undescribed species) rather than against catalogued species alone. The 62 atlas partitions together encompass an estimated 7.3–35.4 M extant species (central ~15.6 M; Supplementary Table S4), so the 494,005 deduplicated eukaryotic standardised labels in the atlas correspond to a ~1.4–6.8% coverage range under this denominator. The width of the range reflects the published spread among global species-richness estimators rather than uncertainty in the atlas itself, and we report it explicitly so the Discussion can cite a single bracket rather than picking one estimator.

Denominators vary in quality and comparability across groups. Vertebrate species counts are relatively stable, while invertebrate and microbial counts are subject to ongoing revision. For prokaryotes, GTDB genome-based species clusters are not directly comparable to traditionally described species, and we note this caveat where relevant. For any group whose partition-canonical tree spans the recognized species pool (Turtles is the prototypical case but the rule is general), the raw tip count is retained in Table 2 but the displayed coverage value is capped at 100%. Groups lacking a plausible or directly comparable denominator are marked with dashes in Table 2 and shown without coverage percentages in the summary figures; a denominator is treated as comparable when it (a) enumerates species in the same operational sense as the tree tips and (b) was last revised within five years of the tree's deposition. Per-tree provenance

for the molecular-tip fractions that underlie the molecular and strict bounds is documented in Supplementary Table S5 (§Per-tree provenance and §Coverage under explicit inclusion criteria).

Crown-age uncertainty extraction

Every dated source tree in the atlas was passed through the audit; the 218 Condamine¹⁴ family-level trees are out of scope for this audit (they are aggregated derivatives, not source chronograms) and are excluded from the 29-tree denominator. For each dated source tree we attempted to recover an estimate of uncertainty on the crown-node age. For BEAST maximum clade credibility (MCC) trees that preserve `[&height_95%HPD={low,high}]` node annotations in the deposited NEXUS file, we parsed all annotated nodes by regular expression and reported the 95% highest posterior density (HPD) interval of the node with the maximum height as the crown-age HPD. For datasets where the authors deposited a posterior sample of trees (10kTrees primates³⁰, 100 trees; Thomson *et al.*³¹ turtles, 100 trees), we parsed each tree, computed its crown age via `max(branching.times())`, and reported the 2.5%/50%/97.5% empirical quantiles across replicates. Empirical quantiles rather than a central HPD are reported here because the posterior and bootstrap distributions are non-parametric and the quantile is the directly comparable analogue of the BEAST 95% HPD. For Stein *et al.*¹⁰ sharks, we used the deposited bootstrap distribution of 500 treePL-dated trees in the same way to produce a 95% bootstrap interval. Of the 29 non-Condamine dated source trees in the atlas, only seven (24%) had a recoverable archival class carrying crown-age uncertainty: three via BEAST MCC HPD annotations (Upham *et al.*⁹ mammals, Nyakatura *et al.*³² carnivores, Vaga *et al.*³³ corals), three via deposited posterior samples (primates, turtles, and Wolfe *et al.*³⁴ decapods), and one via a bootstrap distribution (sharks); the remaining 22 source trees and the 218 Condamine family-level trees were deposited as point estimates with no per-node uncertainty annotation or accompanying posterior or bootstrap distribution. The recoverable archival classes are recorded in the `hpd_method` column of the crown-age inventory (`fig2_hpd_inventory.csv`), which takes the values `beast_mcc_annotation`, `posterior_sample`, or `bootstrap_treepl` for trees with recoverable uncertainty and `none` otherwise; each tree's dating method is recorded separately in the `dating_code` field of Supplementary Table S5. The audit boundary mirrors that of §Per-tree provenance: we did not contact authors, did not re-fit chronograms, and did not search the body of supplementary PDFs for HPDs absent from the deposited tree files. The parsing script is invoked once per rebuild (§Updating cadence and versioning) so a new release-tag re-evaluates uncertainty recovery for any deposited update.

Updating cadence and versioning

The atlas is rebuilt periodically as new phylogenies enter the literature, with no fixed schedule: rebuilds are triggered when the literature-monitoring pipeline of §Literature survey and curation surfaces enough candidate trees to justify a new

release. Each rebuild runs a fixed pipeline: (1) literature monitoring (the survey queries of §Literature survey and curation are rerun against Google Scholar, the listed repositories, the clade-specific databases, and bioRxiv); (2) candidate validation (§Validation); (3) standardization and dictionary regeneration (§Standardization); (4) Supplementary Table S5 regeneration (§Per-tree provenance); (5) partition reassessment (§Partitioning the tree of life); and (6) release tagging on GitHub and deposit on Zenodo.

Partition split/merge events are triggered at step (5) by deterministic rules: a partition *splits* when the rebuild identifies a new clade-bounded comprehensive tree for a sub-clade of an existing partition; two partitions *merge* when a new cross-clade comprehensive tree subsumes both; an unrepresented partition *migrates* into the represented set when a previously absent partition-canonical tree appears. A partition definition does not change for any other reason between releases.

The release scheme is two-level. The atlas-as-resource has a Zenodo *concept-DOI* ([10.5281/zenodo.20127157](https://doi.org/10.5281/zenodo.20127157)) that always resolves to the latest release; each release additionally receives its own Zenodo *version-DOI* (the snapshot reported here is v1.0.8, [10.5281/zenodo.21205025](https://doi.org/10.5281/zenodo.21205025)). GitHub release tags follow semantic versioning (vX.Y.Z) and are aligned one-to-one with the version-DOIs. The repository README points to a single reproduction entry-point notebook that re-derives every number in the paper from the version-pinned standardised trees, dictionary CSV, Supplementary Table S5, and the deposited canonical-succession CSV (`table_s7_canonical_succession.csv`). A GitHub wiki provides human-curatable per-partition review pages, regenerated deterministically at each release from the same data sources (see §Historical succession of canonical trees). An R companion package (`phyloatlas`, submitted to CRAN) provides programmatic access to the standardised trees and per-tree provenance from within R.

To pin a specific snapshot of the atlas, cite the version-DOI; the concept-DOI resolves to the most recent release and is appropriate when citing the atlas as a continuously-maintained resource rather than a fixed deliverable.

Software and AI-assisted tooling

The data-curation pipeline (Catalogue of Life name resolution, per-tree provenance extraction, archival-uncertainty audit), the figure-generation scripts, and the companion `phyloatlas` R package were developed with substantial assistance from large-language-model (LLM) coding tools (Anthropic Claude model family). LLM assistance was used to draft and iteratively refine code, to translate inclusion-criterion rules into reproducible pipeline steps, and to generate exploratory plots; all analytical decisions, scientific judgments, the inclusion or exclusion of individual canonical trees, and the structure of the inclusion-criterion sensitivity envelope are the author's own. All deposited code is open source and inspectable in the project repository, so the LLM contribution to any given script can be audited end-to-end.

Results

A global atlas of empirical phylogenies

We identified and curated 264 trees covering 62 partitions of life across Bacteria, Archaea, and Eukaryota (Table 2, Fig. 1). The 264 trees comprise 218 family-level tetrapod chronograms from Condamine et al. 2019, 25 further partition-canonical phylogenies (one largest available species-level tree per remaining represented partition; together with the Condamine *Crocodylia* tree these constitute the 26 shipped partition-canonicals), and 21 complementary trees: 19 sub-clade phylogenies (e.g., the 5,326-tip Portik et al. frog phylogeny within Amphibians, the 2,258-tip Kawahara et al. butterfly chronogram within Insects) and two cross-cutting reference trees. Each tree was inferred from molecular or genomic data in a peer-reviewed study (or, for two datasets, an established database with published methodology), contains species-level tips, and is freely downloadable from a public repository. The collection was assembled through systematic literature searches across Dryad, Zenodo, Figshare, GitHub, institutional servers, and journal supplements through May 2026.

The largest trees are the Smith & Brown seed plant mega-phylogeny 356,305 tips;¹² the GTDB bacterial reference tree 136,646 genome-based species clusters;²⁹ and the insect-phylo.org insect synthesis 53,596 species;¹⁹. In addition, we include 218 family-level tetrapod trees from Condamine et al. 2019 and the Hedges et al. TimeTree of Life spanning 50,645 species across all domains³⁵.

The representative trees in the current public release were standardized into Newick format with numeric tip labels mapped to a global dictionary of 637,619 unique standardized labels (494,005 eukaryotic standardized labels and 143,614 GTDB prokaryotic genome clusters). This standardization enables cross-tree queries (e.g., finding all trees containing a given species) and lays the groundwork for supertree construction. The current archive includes all representative datasets used for the manuscript, including bryozoa and frogs.

Coverage asymmetry across the tree of life

A central result of our survey is the asymmetry in phylogenetic coverage (Fig. 1). Well-studied clades show near-complete coverage: birds (96.7%), sharks and rays (93.0%; 1,192 of 1,282 described chondrichthyan species in Stein et al. 2018), seed plants (92.9%), mammals (87.8%), ray-finned fish (87.5% under the full Rabosky et al. 2018 TACT-imputed tree; 32.2% under the molecular-only subset shipped in the standardized release, see also Methods), squamates (83.1%), and amphibians (81.7%). For these groups, species-level phylogenies inferred from multi-gene supermatrices have been available for 5–15 years and are routinely used in comparative analyses.

In stark contrast, most invertebrate diversity remains poorly represented in current large-scale phylogenies. Insects, the most species-rich animal class with approximately 1 million described species, are covered by only 53,596 species (5.4%) in the best available tree. Spiders cover 2.8% of described species, crustaceans 0.1%, nematodes 0.6%, and echino-

derms 0.9%. For several major animal phyla, species-level phylogenetic coverage remains sparse or absent. Annelida (~21,000 species) and Platyhelminthes (~20,000 species) lack any broadly representative atlas-scale phylogeny; available molecular work is concentrated in family- or order-level backbones^{36,37}. Bryozoa is represented by the Orr et al. 2022 tree (720 tips overall, of which 517 are cheilostome species; 12.0% of ~6,000 described species are represented in the full tree), and Rotifera (~2,200 species) and 36 additional partitions lack any broadly representative downloadable species-level tree as of May 2026 (Table 2). Mollusca, despite ~85,000 described species, is represented in the atlas only by relatively small backbone trees for gastropods (54 species), cephalopods (81 species), and freshwater bivalves (73 species).

Fungi, despite an estimated 155,000 described species, are represented by only 1,602 species (1.0%) in the most comprehensive genome-scale phylogeny available³⁹. Prokaryotic coverage through GTDB is difficult to compare directly because “species” are defined as genome-based clusters rather than traditional taxonomic descriptions.

The absolute scale of these gaps and the distinction between dated and undated phylogenies are shown in Fig. 1: while seed plants and most vertebrate groups show dark green bars that nearly fill their described-species extent, most invertebrate and microbial groups show only a thin sliver of green against a predominantly black bar, and several major phyla appear as fully black bars because no broadly representative atlas-scale tree is currently available. For rows lacking a directly comparable external denominator, the figure reports tree size but omits a coverage percentage.

Temporal coverage: dated vs. topology-only trees

Of the 264 trees in the atlas, 247 (94%) are time-calibrated (branch lengths in millions of years) and 17 provide topology only or branch lengths in substitutions per site. This is a marked improvement over earlier releases: in 2017 only ~5 dated trees were available. The distinction is critical for macroevolutionary inference: diversification rates, ancestral state reconstruction, and comparative methods all require dated trees.

Time-calibrated trees are available for all vertebrate groups in the collection, most plant datasets (seed plants, ferns, conifers, orchids, Solanaceae, and cacti), four arthropod groups (butterflies, bees, spiders, and crustaceans), and several other datasets (corals, cephalopods, sponges, echinoderms, and diatoms), plus the Condamine family-level trees. The large insect synthesis trees, several invertebrate backbone trees (termites, hemiptera, nematodes, bryozoa, bivalves, and gastropods), the grasses, asteraceae, bryophytes, and eucalyptus trees, as well as the fungal tree and prokaryotic trees, are not time-calibrated, limiting their utility for diversification analyses.

Among the major clade-level trees, crown ages span from 37 Ma (cacti) to 573 Ma (corals), covering much of the Phanerozoic. Where archival uncertainty was recoverable (7 of 29 non-Condamine dated source trees evaluated in the crown-age audit), 95% intervals on crown age span more than an order of magnitude in width, from ~5 Myr for pri-

mates to ~150 Myr for corals (see Methods). Restricting to the 26 partition-canonical source trees, 10 of 26 ship a single point-estimate chronogram, 9 of 26 are undated phylogenomic or supertree topologies, and only 7 of 26 (27%) preserve a posterior or bootstrap distribution from which HPDs can be reconstructed without re-running the dating pipeline. The two ratios enumerate different sets: the crown-age audit (7 of 29) additionally includes sub-clade alternative trees from `table_s6_alternative_trees.csv` with deposited posterior or BEAST MCC HPD annotations, whereas the partition-canonical view (7 of 26) counts only the one canonical tree per partition.

Species accumulation across datasets

Three datasets alone account for ~85% of all species in the collection: seed plants (354,799 unique species after deduplication from 356,305 tips), bacteria (136,646 GTDB clusters), and insects (53,596). The remaining 23 partition-canonical trees and 218 Condamine family-level trees collectively add approximately 97,000 additional species. This concentration means that the overall species count is dominated by a few megadiverse groups, while the “long tail” of smaller trees provides coverage of groups that would otherwise be absent entirely.

Coverage under explicit inclusion criteria

The headline atlas-coverage number, the fraction of described eukaryotic species represented in at least one canonical partition tree, is a function of the inclusion criterion adopted, not a single empirical constant. We report three explicit sensitivity bounds, applied per partition and aggregated across the deduplicated dictionary (Methods). The *permissive* criterion counts every retained tip of the canonical tree, including taxonomically placed and synthesis-grafted tips. The *molecular* criterion removes the non-molecular tips of the three TACT-imputed canonicals (seed plants, squamates, sharks) but otherwise mirrors the permissive criterion. The *strict* criterion further excludes full Open Tree of Life pipeline trees and counts only molecular tips on every remaining partially-imputed canonical. Aggregated across the deduplicated dictionary, the three bounds give 23.7%, ~10%, and ~5.4% of described eukaryotic species respectively, an envelope that spans roughly a factor of four (Table 1).

The per-partition view reveals where the envelope concentrates. Seed plants drift the most under stricter criteria, from 92.9% under the permissive criterion (356,305 tips of Smith & Brown¹² against 383,671 described species in World Flora Online 2025) to ~20.8% under the molecular criterion (and the strict criterion), a ~72-percentage-point shift driven entirely by the Smith & Brown megatree, whose tip pool is overwhelmingly composed of taxonomically grafted species rather than directly molecularly placed ones. Birds show a comparable absolute shift between the permissive/molecular criteria and the strict criterion (96.7% to 0%) because the McTavish et al. 2025 synthesis is a full Open-Tree pipeline product and is excluded outright under the strict criterion. Among the other vertebrate partitions, sharks (93.0% → 47.4%), squamates (83.1% → 35.7%), mammals (87.8% →

60.6% under the strict criterion), and amphibians (81.7% → 54.7% under the strict criterion) all shift visibly under the stricter readings. By contrast, the long tail of small invertebrate, microbial, and protist partitions (spiders, sponges, diatoms, fungi, gastropods, crustaceans) show no drift at all because their canonical trees consist entirely of molecularly placed tips and none are products of the full Open Tree of Life pipeline. The asymmetry between coverage estimate and inclusion criterion is therefore not a uniform multiplicative penalty: it is concentrated on a handful of taxonomically broad, methodologically heterogeneous trees (seed plants, insects, the major tetrapod clades) whose retained tip pool depends sharply on whether one counts grafted or imputed tips.

The Condamine et al. 2019 family-level trees contribute only 1,052 species not already present in the larger trees, indicating that their primary value lies in providing independently estimated family-level topologies for diversification analyses rather than expanding species coverage.

Disaggregating the permissive/molecular/strict envelope by the publication year of each partition’s atlas-shipped canonical reveals when the eukaryotic species pool of the atlas actually accumulated (Fig. 2; derived from `table_s7_canonical_succession.csv`, the deposited canonical-succession ledger). The trajectory is dominated by a single event: Smith & Brown¹² 2018 lifts the permissive count from ~10,000 to ~384,000 species in one step while the molecular and strict counts rise only to ~104,000 and ~100,000 respectively, an illustration of how a single taxonomically-grafted megatree controls the apparent magnitude of the resource, but contributes much less to its strict empirical core. The 2023 step adds the insectphylo v2 collection-backed synthesis (Chesters et al. 2023), which lifts the permissive and molecular counts but is excluded outright under the strict criterion (full-OTL-pipeline tier). By the v1.0.8 snapshot, the three trajectories reach ~464,000 / ~184,000 / ~114,000 species respectively, comparable to (but slightly below) the deduplicated Table 1 headline of 494,005 / ~211,000 / ~114,000; the figure reports summed tip counts pre-deduplication, with the deduplication lift differing across criteria as documented in §Coverage under explicit inclusion criteria.

Overlapping and complementary trees

Several taxonomic groups are covered by multiple independent studies, providing opportunities to assess phylogenetic uncertainty. In the main table, Birds are covered by both the Jetz et al. 2012 posterior distribution (9,993 species × 1,000 trees) and the McTavish et al. 2025 synthesis (10,824 species), and Insects are covered by both the Chesters¹³ species-level tree and the insectphylo.org¹⁹ continuously updated synthesis. The broader collection also contains alternative trees for Mammals two dating methods in⁹, Amphibians (overlapping with the separate Frogs tree), Squamates (Pyron 2013 backbone), and Seed plants (Kew Tree of Life). These overlaps are quantified in Supplementary Table S2.

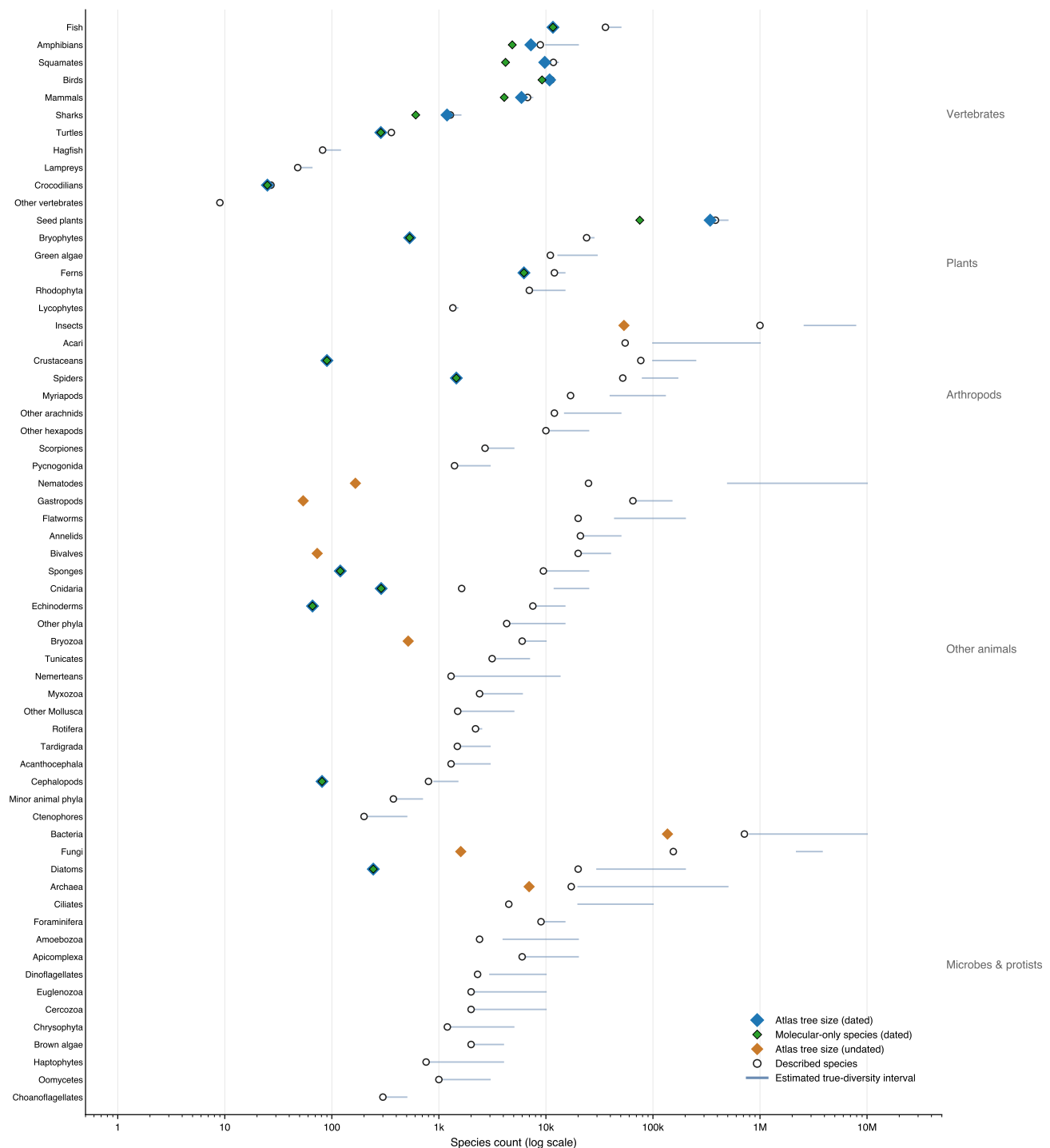


Figure 1 | Phylogenetic coverage across the tree of life (62 partitions). Each row is one partition, grouped by broad taxonomic category. Filled diamonds mark species counts in the canonical atlas tree for that partition: **blue diamond**: total tip count of the canonical time-calibrated tree (all tips, including taxonomically placed and synthesis-grafted species); **green diamond**: molecular-only tip count of the same time-calibrated tree (tips with direct molecular sequence support, excluding taxonomic placements and grafted tips); **amber diamond**: tip count of the canonical tree where it is undated. Open circles mark described species counts from authoritative taxonomic catalogues; horizontal lines span published estimated true-diversity intervals (low–high). The horizontal axis is log species count. For dated rows whose canonical tree has direct molecular support for every tip, the blue and green diamonds coincide and appear as a single marker; for dated rows with partial taxonomic placement (e.g., seed plants, squamates, mammals, birds, sharks), the green diamond sits visibly to the left of the blue, with the horizontal gap indicating how much of the tree’s tip pool is taxonomically rather than molecularly placed. Vertebrate and plant partitions cluster on the right of the plot (high described counts, near-complete trees); microbe and minor-animal-phyla partitions show open circles without diamonds, indicating no atlas-scale phylogeny is currently available. Broad taxonomic categories are labelled on the right margin.

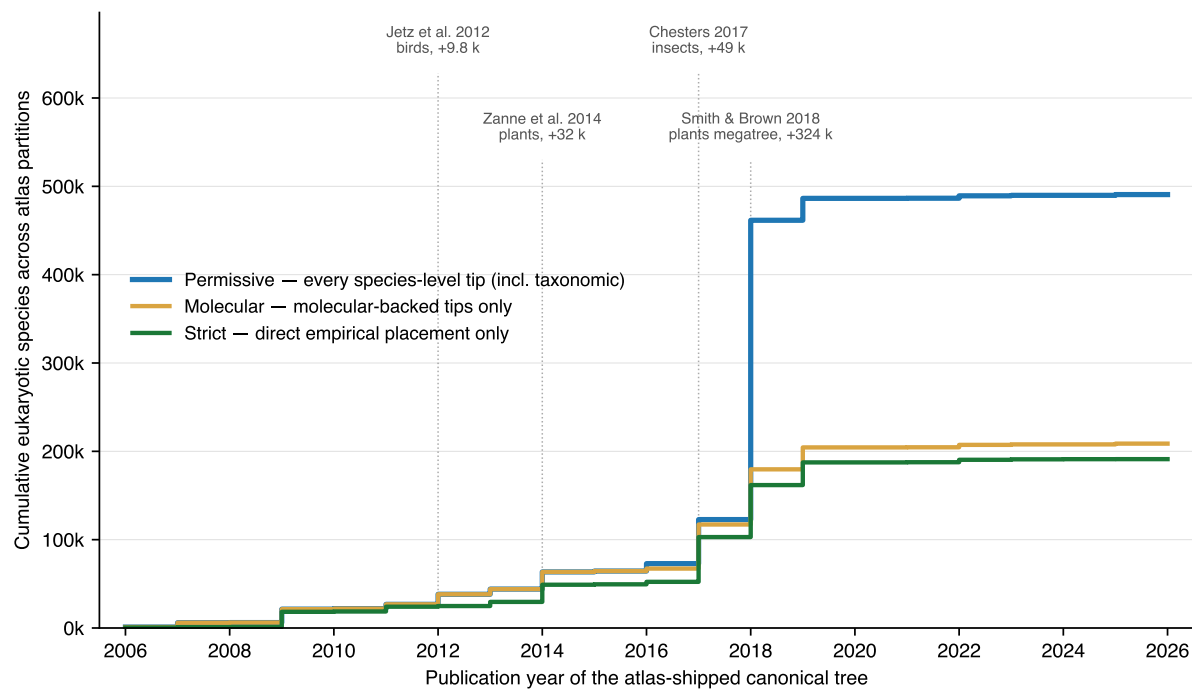


Figure 2 | Cumulative eukaryotic species in the atlas through time, under three inclusion criteria. For each year between 2006 and 2025 we take, per partition and per criterion, the running maximum of species contribution across all canonical trees from Supplementary Table S7 published at or before that year, then sum across partitions. The running-max semantics reflects when species first became available to the community, not when the atlas project later adopted a particular tree, so e.g. the bird partition's contribution accumulates with Jetz et al.²³ 2012 (+ ~9.8 k species, the first comprehensive bird tree), not in 2025 when the atlas adopted McTavish et al.¹¹ (an additional + ~0.8 k species refinement). **Blue (permissive):** all retained tips of the running-max canonical, including taxonomically placed and synthesis-grafted tips. **Amber (molecular):** the three TACT-imputed canonicals (seed plants, squamates, sharks) contribute only their molecular-tip fraction; other canonicals contribute in full. **Green (strict):** rows whose `otl_pipeline_involvement` is full-pipeline or collection-backed contribute zero; other rows with molecular-tip fraction < 1 contribute molecular tips only. Four highlighted increments label the largest historical step-changes: Jetz 2012 birds, Zanne 2014 plants, Chesters 2017 insects, and Smith & Brown 2018's plant megatree (the dominant single event, + ~324 k tips of which ~22.6% are molecular). By the v1.0.8 snapshot the three trajectories reach ~491,000 / ~209,000 / ~191,000 species respectively. The permissive and molecular endpoints align with Table 1's deduplicated headline (494,005 / ~211,000) within the name-resolution lift documented in Methods. The Figure-2 strict endpoint (~191,000) exceeds Table 1's strict headline (~114,000) because the figure uses row-level OTL-pipeline exclusion (Chesters 2020's 69,000-tip non-OTL insect tree contributes under the strict criterion even though the current atlas-shipped insect canonical, Chesters 2023, is excluded) whereas Table 1 reports the stricter partition-level reading (birds and insects entirely excluded). Both readings are reported here so readers can choose the bound appropriate to their analysis. Prokaryotic partitions (bacteria, archaea) are excluded to keep the y-axis comparable with the eukaryotic-only permissive/molecular/strict framework.

Discussion

To our knowledge, this is the first broad cross-kingdom quantitative synthesis of species-level phylogenetic coverage in currently downloadable empirical phylogenies. The central finding (that only 23.7% of described eukaryotic species are represented in the largest currently available empirical phylogenies, with 36 atlas partitions lacking any broadly representative downloadable tree) has immediate implications for systematic biology, comparative methods, and biodiversity science.

Coverage asymmetry: observations and plausible drivers

The coverage asymmetry we document is the most conspicuous feature of the current empirical tree of life. A formal causal analysis is beyond the scope of this resource paper,

but several factors plausibly contribute. Vertebrate systematics has benefited from sustained museum infrastructure and comprehensive specimen archives⁴⁰; conservation agencies (IUCN, CITES) have driven comprehensive species inventories for vertebrates that subsequently motivated phylogenetic efforts. Scale is also a straightforward constraint: building a phylogeny for ~6,700 mammals is a tractable decade-long project, whereas building one for 85,000 mollusks or 1,000,000 insects requires fundamentally different approaches⁴¹. Universal vertebrate molecular markers (e.g., cytochrome b, RAG1) enabled the supermatrix approach that produced the large vertebrate trees^{9,18}, while invertebrate groups more often require order-specific primer development, though phylogenomic approaches are increasingly circumventing this limitation. Finally, methodological choices within groups shape apparent coverage: birds reach 96.7% in McTavish et al. 2025 2025, a synthesis tree that integrates

Jetz et al. 2012's taxonomic-placement framework with Open Tree topology to assign essentially every recognized species (Jetz 2012 alone covers 89.3%), whereas the mammal supertree⁹ did not impute all data-deficient species, leaving coverage at 87.8% despite comparable taxonomic completeness. Coverage therefore reflects not only biological tractability but also the ambition and methodology of each tree-building effort. A more comprehensive discussion of institutional, funding, and community-level drivers is provided in the Supplementary Discussion.

The time-calibration gap

The absence of time-calibrated trees for the largest invertebrate and microbial clades represents a second critical gap for systematic biology. Without dated phylogenies, researchers cannot estimate diversification rates, test hypotheses about the tempo and mode of evolution, or place evolutionary events in geological context. Out of 264 atlas trees, 247 are dated, but the undated subset includes the most consequential synthesis resources: the insectphylo.org and Chesters insect supertrees (53,596 and 49,338 species respectively), the kingdom-Fungi tree (1,602 species), and the GTDB bacterial + archaeal trees (143,614 genome clusters); the undated set also includes the canonical phylogenomic ML trees for bivalves (Pfeiffer *et al.*⁴²), bryozoa (Orr *et al.*³⁸), and gastropods (Zapata *et al.*⁴³), which are deposited as ML/Bayesian topologies without time calibration. Time-calibrating these existing trees should be a priority for the field. Among the trees that *are* dated, the divergence-time method itself varies widely, from Bayesian relaxed clocks to re-dated supertrees, with correspondingly different capacities to preserve age uncertainty (Fig. 3).

Statistical consequences of discarding divergence-time uncertainty

A second under-archival pattern complements the time-calibration gap. Each deposited dated tree summarizes a posterior or bootstrap distribution that, in principle, encodes node-level uncertainty, typically reported as 95% HPD intervals on every internal node in the original publication. In practice, however, this uncertainty is rarely propagated through to the deposited derivative tree. Of the 29 non-Condamine dated source trees in our atlas, only seven (24%) preserve recoverable archival uncertainty on the crown age: three through BEAST MCC trees with embedded `height_95%_HPD` node annotations, three through deposited posterior samples of trees, and one through a bootstrap-replicate distribution (Methods). The remaining 22 source trees and all 218 Condamine¹⁴ family-level trees are deposited as point-estimate chronograms with no per-node uncertainty, even where the original inference (Bayesian or bootstrapped) clearly produced one. Where uncertainty is recoverable, interval widths span more than an order of magnitude, from ~5 Myr for well-calibrated primates to ~150 Myr for corals, reflecting heterogeneous calibration density that downstream users of the trees cannot otherwise see. The systematic loss of uncertainty information between inference and archival is itself a methodological gap with direct consequences: comparative analyses that propagate phylogenetic uncertainty across the tree of life cannot do so

when only point estimates are deposited.

The consequences are statistical, not merely bookkeeping. A downstream comparative analysis treats each node or crown age as an input; when only a point estimate is deposited, that input is silently promoted from a random variable to a fixed constant, with three distinct effects. First, *variance is underestimated*: a statistic computed on a single plugged-in chronogram (a net-diversification rate, a phylogenetically independent contrast, or the slope of an age-dependent regression) inherits none of the calibration uncertainty, so its standard error is anticonservative and its interval too narrow, inflating Type I error in precisely the age-dependent tests the tree was built to support. Second, point estimates are *biased*, not merely over-confident, whenever the downstream quantity is nonlinear in the ages: a per-lineage rate scales as (number of events)/ t , and by Jensen's inequality $E[1/t] \neq 1/E[t]$, so evaluating the estimator at the mean age does not return the mean rate; the target expectation is unrecoverable once the age distribution is discarded. Third, the *archival class determines which stratum of uncertainty survives*, and the strata are not interchangeable; each retains a strictly weaker summary than the last. A deposited posterior sample of trees retains the full *joint* distribution of node ages, including the across-node covariances on which tree-wide methods such as independent contrasts and phylogenetic generalized least squares depend. Per-node 95% HPD annotations on a single summary tree retain only each node's *marginal*, discarding those covariances. A bootstrap-replicate set encodes a frequentist sampling distribution rather than a posterior. A bare point estimate retains nothing. The archival class (recorded in `hpd_method`) is thus also a ladder of statistical usability (posterior sample > MCC-with-HPD > bootstrap distribution > point estimate) on which the dating *method* sets the ceiling (only a Bayesian relaxed clock can supply the full joint posterior; penalised-likelihood, supertree, and synthesis chronograms cannot) while the deposit format fixes where a tree actually lands. Collapsing a Bayesian posterior to a point-estimate deposit therefore discards recoverable, inference-load-bearing information, not a nuisance annotation.

Routine archival of MCC trees with full per-node HPD annotation, or of at least a moderately sized posterior subsample, would address this gap at negligible storage cost.

Limitations

This collection is a curated atlas of the largest publicly downloadable empirical phylogenies currently available for major clades (comprising 264 trees across 62 atlas partitions), not an exhaustive census of every published phylogenetic placement in the literature. Many species appear in smaller clade-specific or regional trees that were not included here because our search prioritized the most comprehensive available resource for each group. Coverage estimates are therefore best interpreted as representation in current large-scale phylogenetic resources, and the true fraction of species with *some* empirical phylogenetic placement is likely higher than 23%.

Comparisons across clades are approximate because taxonomic practice, species delimitations, denominator sources,



Figure 3 | Dating-method provenance across the atlas. Each marker is a dated clade at its crown age (x) and observed tip count (y); the vertical band spans the described-to-estimated species-richness range $[N_d, N_u]$. Colour encodes the divergence-time *dating method* of the tree's original source: Bayesian relaxed clock (BRC; BEAST), penalised likelihood (PL; TREEPL), dated synthesis (SYN) and supertree (ST). All 246 clades are method-coded (BRC 144, PL 33, ST 68, SYN 1). The 218 Condamine et al. (2019) compiled family trees are split by their inherited source method (131 BRC via BEAST; 66 ST; 21 PL via TREEPL) rather than lumped as one "Condamine" category, which would be a compiler label, not a method. Method definitions and the orthogonal archival-class (*hpd_method*) analysis are given in the Supplement.

and data availability differ substantially among groups. Prokaryotic “species” defined by genome-similarity thresholds in GTDB are not directly equivalent to traditionally described eukaryotic species. Invertebrate species counts are under active revision, and some denominators may be substantially revised in coming years. Additionally, our search was limited to English-language literature and publicly accessible repositories; trees deposited in regional databases or behind journal paywalls may have been missed.

The scale of the gap in context

The 62 partitions in our atlas collectively encompass ~2.9 million described species and an estimated 7.3–35.4 million species worldwide (central ~15.6 million) (Fig. 1; Supplementary Table S4). Published estimates of global species richness vary by orders of magnitude depending on scope and methodology (Supplementary Table S7). For eukaryotes, the most widely cited figure is Mora et al.’s⁴⁴ ~8.7 million (± 1.3 M), while Costello et al. 2013 argue for a lower ~5 million after correcting for synonymy. For prokaryotes, estimates range from ~1 million⁴⁶ to $\sim 10^{12}$ ⁴⁷. When host-specific symbionts and cryptic species are included, Larsen et al. 2017 estimate 1–6 billion species across all life. Our figure of ~15.6 million sits between the eukaryote-only consensus and the more expansive all-life estimates, reflecting our inclusion of prokaryotic genome clusters and conservative per-group estimates from authoritative sources rather than extrapolation from scaling laws. Regardless of which global total one adopts, the phylogenetic coverage documented here (637,619 standardized labels in empirical trees) represents a small fraction of estimated global diversity. This shows both the scale of the gap and the value of a standardized atlas that makes the existing coverage transparent.

Toward a comprehensive empirical tree of life

Our standardized collection (with 637,619 unique labels mapped through a global dictionary of numeric IDs) provides a foundation for several immediate applications in systematic biology: cross-tree species lookup (identifying all trees containing a given species), overlap quantification between independent studies, and benchmarking of coverage gaps across clades. The interoperable format is also suited to future supertree or megatree construction. Together with the Open Tree of Life taxonomy as a scaffold, these empirical phylogenies could be integrated into a more comprehensive empirical tree of life. The cumulative-coverage trajectory in Fig. 2, from ~1,000 species in 2006 to ~487,000 today under the permissive criterion, suggests that the remaining unrepresented partitions are tractable on a decade timescale rather than a distant one.

Acknowledgements

We thank the authors of all phylogenetic studies included in this collection for making their data publicly available. The data-curation pipeline, figure-generation scripts, and the companion *phyloatlas* R package were developed with substantial assistance from large-language-model coding tools

(Anthropic Claude model family), as described in Materials and Methods; all analytical decisions and scientific judgments are the author’s own.

Funding

This research was supported through the research programme “Statistical Frontiers in Dynamic Network Modeling,” funded by the Swiss National Science Foundation (project no. FNS 240065), directed by Prof. Ernst-Jan Camiel Wit.

Author contributions

F.R. conceived the study, assembled and curated the dataset, performed all analyses, and wrote the manuscript.

Competing interests

The author declares no competing interests.

Data availability

The complete collection is available at <https://github.com/franciscorichter/phylo-species-atlas> and archived on Zenodo under the concept DOI <https://doi.org/10.5281/zenodo.20127157>¹. Per-partition human-curatable review pages are maintained as a GitHub wiki at <https://github.com/franciscorichter/phylo-species-atlas/wiki>, with one page per atlas partition synthesizing the canonical tree, the historical succession (Supplementary Table S7), the audit status, and a community-editable notes section. An R client package (*phyloatlas*) for programmatic access is available at <https://github.com/franciscorichter/phylo-species-atlas/tree/main/phyloatlas> (submitted to CRAN, 2026-05-31). The standardized trees (Newick format with numeric labels), global species dictionary, full label mapping, and metadata are provided as a version-controlled, evolving resource that will be updated as improved phylogenies become available. All underlying trees are permanently archived in their original repositories (Dryad, Figshare, Zenodo, GitHub) as cited; this collection provides standardized derivatives with tip-label normalization and cross-tree interoperability. The synthetic TimeTree-of-Life chronogram³⁵ is included as a cited cross-cutting reference but is not redistributed in this collection, at the request of the TimeTree project; it is available directly from <https://timetree.org>.

Code availability

All curation, standardization, and analysis scripts are available at <https://github.com/franciscorichter/phylo-species-atlas>.

Additional information

Supplementary information Supplementary material accompanies this bioRxiv preprint.

¹The concept DOI always resolves to the latest version; readers requiring a static snapshot should pin to the specific version DOI (the release reported here is v1.0.8: <https://doi.org/10.5281/zenodo.21205025>).

Correspondence and requests for materials should be addressed to Francisco Richter.

References

- [1] Hélène Morlon. Phylogenetic approaches for studying diversification. *Ecology Letters*, 17:508–525, 2014.
- [2] Nick JB Isaac, Samuel T Turvey, Ben Collen, Carly Waterman, and Jonathan EM Baillie. Mammals on the EDGE: conservation priorities based on threat and phylogeny. *PLOS ONE*, 2:e296, 2007.
- [3] Jeffrey B Joy, Richard H Liang, Rosemary M McCloskey, T Nguyen, and Art FY Poon. Ancestral reconstruction. *PLOS Computational Biology*, 12:e1004763, 2016.
- [4] Fredrik Ronquist and Isabel Sanmartín. Phylogenetic methods in biogeography. *Annual Review of Ecology, Evolution, and Systematics*, 42:441–464, 2011.
- [5] Paschalia Kapli, Ziheng Yang, and Maximilian J Telford. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics*, 21(7):428–444, 2020.
- [6] Cody E Hinchliff, Stephen A Smith, James F Allman, J Gordon Burleigh, Ruchi Chaudhary, Lyndon M Coghill, Keith A Crandall, Jiabin Deng, Bryan T Drew, Romina Gazis, et al. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences*, 112(41):12764–12769, 2015.
- [7] William H Piel, Michael J Donoghue, and Michael J Sanderson. TreeBASE v.2: A Database of Phylogenetic Knowledge. *e-BioSphere*, 2009.
- [8] Y Jiang et al. TreeHub: a comprehensive dataset of phylogenetic trees. *Scientific Data*, 2025. doi: 10.1038/s41597-025-05282-4.
- [9] Nathan S Upham, Jacob A Esselstyn, and Walter Jetz. Inferring the mammal tree: species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLOS Biology*, 17(12):e3000494, 2019.
- [10] Rachel W Stein, Christopher G Mull, Tyler S Kuhn, Neil C Aschliman, Lindsay NK Davidson, Jeffrey B Joy, Graham J Smith, Nicholas K Dulvy, and Arne O Mooers. Global priorities for conserving the evolutionary history of sharks, rays and chimaeras. *Nature Ecology & Evolution*, 2:288–298, 2018.
- [11] Emily Jane McTavish et al. A complete and dynamic tree of birds. *Proceedings of the National Academy of Sciences*, 2025. doi: 10.1073/pnas.2409658122.
- [12] Stephen A Smith and Joseph W Brown. Constructing a broadly inclusive seed plant phylogeny. *American Journal of Botany*, 105(3):302–314, 2018.
- [13] Douglas Chesters. Construction of a species-level tree of life for the insects and utility in taxonomic profiling. *Systematic Biology*, 66(3):426–439, 2017.
- [14] Fabien L Condamine, Jonathan Rolland, and Hélène Morlon. Assessing the causes of diversification slowdowns: temperature-dependent and diversity-dependent models receive equivalent support. *Ecology Letters*, 22:1900–1912, 2019.
- [15] Emmanuel Paradis and Klaus Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3):526–528, 2019.
- [16] Catalogue of Life Partnership. Catalogue of life. <https://www.catalogueoflife.org/>, 2025. Accessed May 2026.
- [17] João Filipe Riva Tonini, Karen H Beard, Rodrigo B Ferreira, Walter Jetz, and R Alexander Pyron. Fully-sampled phylogenies of squamates reveal evolutionary patterns in threat status. *Biological Conservation*, 204:23–31, 2016.
- [18] Daniel L Rabosky, Jonathan Chang, Pascal O Title, Peter F Cowman, Lauren Sallan, Matt Friedman, Kristin Kaschner, Cristina Garilao, Thomas J Near, Marta Coll, and Michael E Alfaro. An inverse latitudinal gradient in speciation rate for marine fishes. *Nature*, 559:392–395, 2018.
- [19] Douglas Chesters et al. Launching insectphylo.org: a new hub facilitating construction and use of synthesis molecular phylogenies of insects. *Molecular Ecology Resources*, 2023. doi: 10.1111/1755-0998.13817.
- [20] Walter Jetz and R Alexander Pyron. The interplay of past diversification and evolutionary isolation with present imperilment across the amphibian tree of life. *Nature Ecology & Evolution*, 2:850–858, 2018.
- [21] Daniel M Portik, Jeffrey W Streicher, and John J Wiens. Frog phylogeny: a time-calibrated, species-level tree based on hundreds of loci and 5,242 species. *Molecular Phylogenetics and Evolution*, 188:107907, 2023.
- [22] Akito Y Kawahara et al. A global phylogeny of butterflies reveals their evolutionary history, ancestral hosts and biogeographic origins. *Nature Ecology & Evolution*, 7:903–913, 2023.
- [23] Walter Jetz, Gavin H Thomas, Jeffrey B Joy, Klaas Hartmann, and Arne O Mooers. The global diversity of birds in space and time. *Nature*, 491:444–448, 2012.
- [24] R. Alexander Pyron, Frank T. Burbrink, and John J. Wiens. A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. *BMC Evolutionary Biology*, 13:93, 2013. doi: 10.1186/1471-2148-13-93.
- [25] William J. Baker, Paul Bailey, Vanessa Barber, Abigail Barker, Sidonie Bellot, David Bishop, Laura R. Botigué, Grace Brewer, Tom Carruthers, James J. Clarkson, Jeffrey Cook, Robyn S. Cowan, Steven Dodsworth, Niroshini Eritawale, Elaine François, Berta Gallego, Matthew G. Johnson, Jan T. Kim, Kevin Leempoel, Olivier Maurin, Catherine McGinnie, Lisa Pokorny, Shyamali Roy, Malcolm Stone, Eduardo Toledo, Norman J. Wickett, Alexandre R. Zuntini, Wolf L. Eisehard, Paul J. Kersey, Ilia J. Leitch, and Félix Forest. A comprehensive phylogenomic platform for exploring the angiosperm tree of life. *Systematic Biology*, 71(2):301–319, 2022. doi: 10.1093/sysbio/syab035.
- [26] Connor J Burgin, Jocelyn P Colella, Philip L Kahn, and Nathan S Upham. How many species of mammals are there? *Journal of Mammalogy*, 99(1):1–14, 2018.
- [27] Peter Uetz, Paul Freed, and Jiří Hošek. The Reptile Database. <http://www.reptile-database.org>, 2025. Accessed May 2026.
- [28] Darrel R Frost. Amphibian species of the world: An online reference. version 6.2. <https://amphibiansoftheworld.amnh.org/>, 2025. American Museum of Natural History, New York. Accessed May 2026.
- [29] Donovan H Parks et al. GTDB release 10: a complete and systematic taxonomy for 715,230 bacterial and 17,245 archaeal genomes. *Nucleic Acids Research*, 2025. doi: 10.1093/nar/gkaf1040.
- [30] Christian Arnold, Luke J Matthews, and Charles L Nunn. The 10kTrees website: a new online resource for primate phylogeny. *Evolutionary Anthropology*, 19:114–118, 2010.
- [31] Robert C Thomson, Phillip Q Spinks, and H Bradley Shaffer. A global phylogeny of turtles reveals a burst of climate-associated diversification on continental margins. *Proceedings of the National Academy of Sciences*, 118(7):e2012215118, 2021.

- [32] Katrin Nyakatura and Olaf R P Bininda-Emonds. Updating the evolutionary history of Carnivora (Mammalia): a new species-level supertree complete with divergence time estimates. *BMC Biology*, 10:12, 2012.
- [33] Claudia Francesca Vaga, Andrea M Quattrini, et al. A global coral phylogeny reveals resilience and vulnerability through deep time. *Nature*, 2025. doi: 10.1038/s41586-025-09615-6.
- [34] Joanna M Wolfe et al. A phylogenomic framework, evolutionary timeline and genomic resources for comparative studies of decapod crustaceans. *Proceedings of the Royal Society B*, 286: 20190079, 2019.
- [35] S Blair Hedges, Julie Marin, Michael Suleski, Madeline Paymer, and Sudhir Kumar. Tree of life reveals clock-like speciation and diversification. *Molecular Biology and Evolution*, 32(4):835–845, 2015.
- [36] Anne Weigert, Conrad Helm, Matthias Meyer, Birgit Nickel, Detlev Arendt, Bernhard Hausdorf, Scott R Santos, Kenneth M Halanych, Günter Purschke, Christoph Blömer, and Torsten H Struck. Illuminating the base of the annelid tree using transcriptomics. *Molecular Biology and Evolution*, 31(6):1391–1401, 2014.
- [37] Jan Brabec et al. Molecular phylogeny of the Platyhelminthes reveals an early, rapid diversification of Neodermata and the evolution of parasitism. *Current Biology*, 33(21):4507–4514, 2023.
- [38] Russell J S Orr, Emanuela Di Martino, Mali H Ramsfjell, Dennis P Gordon, Björn Berning, Lee Hsiang Liow, et al. Paleozoic origins of cheilostome bryozoans and their parental care inferred by a new genome-skimmed phylogeny. *Science Advances*, 8(12):abm7452, 2022. doi: 10.1126/sciadv.abm7452.
- [39] Yuanning Li et al. A genome-scale phylogeny of the kingdom Fungi. *Current Biology*, 31:1653–1665, 2021.
- [40] Adrian M Lister and Climate Change Research Group. Natural history collections as sources of long-term datasets. *Trends in Ecology & Evolution*, 26(4):153–154, 2011.
- [41] Nigel E Stork. How many species of insects and other terrestrial arthropods are there on Earth? *Annual Review of Entomology*, 63:31–45, 2018.
- [42] John M Pfeiffer, Jesse W Breinholt, and Lawrence M Page. Unioverse: a phylogenomic resource for reconstructing the evolution of freshwater mussels (Bivalvia, Unionoida). *Molecular Phylogenetics and Evolution*, 137:114–126, 2019.
- [43] Felipe Zapata et al. Phylogenomic analyses of deep gastropod relationships reject Orthogastropoda. *Proceedings of the Royal Society B*, 281:20141739, 2014.
- [44] Camilo Mora, Derek P Tittensor, Sina Adl, Alastair G B Simpson, and Boris Worm. How many species are there on Earth and in the ocean? *PLoS Biology*, 9(8):e1001127, 2011.
- [45] Mark J Costello, Robert M May, and Nigel E Stork. Can we name Earth’s species before they go extinct? *Science*, 339(6117):413–416, 2013.
- [46] Stilianos Louca, Florent Mazel, Michael Doebeli, and Laura Wegener Parfrey. A census-based estimate of Earth’s bacterial and archaeal diversity. *PLoS Biology*, 17(2):e3000106, 2019.
- [47] Kenneth J Locey and Jay T Lennon. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, 113(21):5970–5975, 2016.
- [48] Brendan B Larsen, Elizabeth C Miller, Matthew K Rhodes, and John J Wiens. Inordinate fondness multiplied and redistributed: the number of species on Earth and the new pie of life. *The Quarterly Review of Biology*, 92(3):229–265, 2017.
- [49] Alexander A Stewart and John J Wiens. A time-calibrated salamander phylogeny including 765 species and 503 genes. *Molecular Phylogenetics and Evolution*, 204:108272, 2025.
- [50] Victor A Tagliacollo et al. A time-calibrated backbone phylogeny for Neotropical freshwater fishes. *Frontiers in Bioinformatics*, 2024. doi: 10.3389/fbinf.2024.1433995.
- [51] Kevin R Burgio et al. Phylogenetic supertree and functional trait database for all extant parrots. *Data in Brief*, 24:103882, 2019.
- [52] Michael R McGowen, Georgia Tsagkogeorga, Sandra Álvarez-Carretero, Mario dos Reis, Matthew Struebig, Robert Deaville, Paul D Jepson, Simon Jarman, Andrea Polanowski, Phillip A Morin, and Stephen J Rossiter. Phylogenomic resolution of the cetacean tree of life using target sequence capture. *Systematic Biology*, 69(3):479–501, 2020.
- [53] Patricia A Henríquez-Piskulich et al. A supermatrix phylogeny of the world’s bees (Hymenoptera: Anthophila). *Molecular Phylogenetics and Evolution*, 190:107963, 2024.
- [54] Jonathan Romiguier et al. Ant phylogenomics reveals a natural selection hotspot preceding the origin of complex eusociality. *Current Biology*, 32:2942–2947, 2022.
- [55] Kevin P Johnson et al. Phylogenomics and the evolution of hemipteroid insects. *Proceedings of the National Academy of Sciences*, 115(50):12775–12780, 2018.
- [56] Simon Hellemans et al. Genomic data provide insights into the classification of extant termites. *Nature Communications*, 15:6712, 2024.
- [57] Nicole L Garrison et al. Spider phylogenomics: untangling the Spider Tree of Life. *PeerJ*, 4:e1719, 2016.
- [58] Joel H Nitta, Eric Schuettelpelz, Santiago Ramírez-Barahona, and Wataru Iwasaki. An open and continuously updated fern tree of life. *Frontiers in Plant Science*, 13:909768, 2022.
- [59] Jin-Hua Ran, Ting-Ting Shen, Hui Wu, Xun Gong, and Xiao-Quan Wang. Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. *Molecular Phylogenetics and Evolution*, 129:106–116, 2018.
- [60] Julia Bechteler et al. Comprehensive phylogenomic time tree of bryophytes reveals deep relationships and uncovers gene incongruences in the last 500 million years of diversification. *American Journal of Botany*, 110:e16249, 2023.
- [61] Grass Phylogeny Working Group III. A nuclear phylogenomic tree of grasses (Poaceae) recovers current classification despite gene tree incongruence. *New Phytologist*, 245:818–834, 2025. doi: 10.1111/nph.20263.
- [62] Jamie B Thompson, Katie E Davis, Harry O Dodd, Matthew A Wills, and Nicholas K Priest. Speciation across the Earth driven by global cooling in terrestrial orchids. *Proceedings of the National Academy of Sciences*, 120:e2102408120, 2023.
- [63] Jennifer R Mandel et al. A fully resolved backbone phylogeny reveals numerous dispersals and explosive diversifications throughout the history of Asteraceae. *Proceedings of the National Academy of Sciences*, 116:14083–14088, 2019.
- [64] Tiina Särkinen, Lynn Bohs, Richard G Olmstead, and Sandra Knapp. A phylogenetic framework for evolutionary study of the nightshades (Solanaceae): a dated 1000-tip tree. *BMC Evolutionary Biology*, 13:214, 2013.
- [65] Jamie B Thompson et al. CactEcoDB: a comprehensive ecological database for cacti. *Scientific Data*, 2026. doi: 10.1038/s41597-026-06936-7.
- [66] Michael D Crisp et al. Perianth evolution and implications for generic delimitation in the eucalypts (Myrtaceae), including the

- description of the new genus, *Blakella*. *Journal of Systematics and Evolution*, 62:895–916, 2024. doi: 10.1111/jse.13047.
- [67] Kiran Basava et al. A phylogeny of extant coleoid cephalopods with brain data. *bioRxiv*, 2024. doi: 10.1101/2024.04.29.591691.
- [68] Xue Qing et al. Phylogenomic insights into the evolution and origin of Nematoda. *Systematic Biology*, 74:349–363, 2025.
- [69] Dennis V Lavrov et al. Phylomitogenomics bolsters the high-level classification of Demospongiae (phylum Porifera). *PLOS ONE*, 2023.
- [70] Nicolás Mongiardino Koch and Jeffrey R Thompson. Phylogenomic analyses of echinoid diversification prompt a re-evaluation of their fossil record. *eLife*, 11:e72460, 2022.
- [71] Teofil Nakov, Jeremy M Beaulieu, and Andrew J Alverson. Accelerated diversification is related to life history and locomotion in a hyperdiverse lineage of microbial eukaryotes (Diatoms, Bacillariophyta). *New Phytologist*, 219(1):462–473, 2018. doi: 10.1111/nph.15137.
- [72] Kelsey Williamson et al. A robustly rooted tree of eukaryotes reveals their excavate ancestry. *Nature*, 640:974–981, 2025.

Table 2 | The PhyloDatasets collection. Each row represents one dataset (source study or established database). Columns: taxonomic group, primary study, publication year, journal, number of species-level tips in the representative tree, estimated total described species for that group, coverage (%), whether the tree is time-calibrated, and the source for described species estimates. Groups covered by multiple studies appear on separate rows. The final section lists major taxonomic groups ($\geq 1,000$ described species) for which no broadly representative downloadable species-level phylogeny was identified in our survey, plus an aggregate row for 21 additional minor phyla (see footnote ¹). Dashes indicate data not directly comparable or not available. Per-dataset details including DOIs and download URLs are maintained at the live atlas website (<https://francisrichter.github.io/phylo-species-atlas/>).

Group	Study	Year	Journal	Tips	Desc.	%	Dated	Sp. count	source
Vertebrates									
Mammals	Upham et al. ⁹	2019	PLOS Biol	5,912	6,736	87.8	Yes	MDD	
Frogs	Portik et al. ²¹	2023	MPE	5,326	7,486	71.1	Yes	AmphibiaWeb	
Salamanders	Stewart and Wiens ⁴⁹	2025	MPE	796	814	97.8	Yes	AmphibiaWeb	
Fish	Rabosky et al. ¹⁸	2018	Nature	11,638	36,000	32.2	Yes	Eschmeyer's	
Neotrop. fish	Tagliacollo et al. ⁵⁰	2024	Front Bioinf	3,167	6,200	51.1	Yes	—	
Turtles	Thomson et al. ³¹	2021	PNAS	274	360	76.1	Yes	Reptile DB	
Primates	Arnold et al. ³⁰	2010	Evol Anthro	272	522	52.1	Yes	IUCN	
Parrots	Burgio et al. ⁵¹	2019	Data Brief	413	413	100	Yes	IUCN	
Carnivora	Nyakatura and Bininda-Emonds ³²	2012	BMC Biol	294	294	100	Yes	IUCN	
Cetaceans	McGowen et al. ⁵²	2020	Syst Biol	85	92	92.4	Yes	WoRMS	
Tetrapod fam.	Condamine et al. ¹⁴	2019	Ecol Lett	25	27	92.6	Yes	IUCN CSG	
Arthropods									
Insects	insectphylo.org ¹⁹	2023	Mol Ecol Res	53,596	1,000,000	5.4	No	CoL	
Insects	Chesters ¹³	2017	Syst Biol	49,338	1,000,000	4.9	No	CoL	
Butterflies	Kawahara et al. ²²	2023	Nat Ecol Evol	2,258	19,000	11.9	Yes	CoL	
Bees	Henriquez-Piskulich et al. ⁵³	2024	MPE	4,586	20,507	22.4	Yes	Discover Life	
Ants	Romiguier et al. ⁵⁴	2022	Curr Biol	83	14,415	0.6	No	AntCat	
Hemiptera	Johnson et al. ⁵⁵	2018	PNAS	193	95,000	0.2	No	—	
Termites	Hellemans et al. ⁵⁶	2024	Nat Comms	138	3,000	4.6	No	—	
Spiders	Garrison et al. ⁵⁷	2016	PeerJ	1,456	52,249	2.8	Yes	WSC	
Crustaceans	Wolfe et al. ³⁴	2019	Proc R Soc B	90	77,000	0.1	Yes	WoRMS	
Plants									
Ferns	Nitta et al. ⁵⁸	2022	Front Plant Sci	6,234	12,000	52.0	Yes	PPG I	
Conifers	Ran et al. ⁵⁹	2018	MPE	16 ^b	615	2.6	Yes	—	
Bryophytes	Bechteler et al. ⁶⁰	2023	Am J Bot	533	24,000	2.2	No	CoL	
Grasses	GPWG Grass Phylogeny Working Group III ⁶¹	2025	New Phytol	1,153	12,000	9.6	No	CoL	
Orchids	Thompson et al. ⁶²	2023	PNAS	1,475	28,000	5.3	Yes	CoL	
Asteraceae	Mandel et al. ⁶³	2019	PNAS	256	34,000	0.8	No	CoL	
Solanaceae	Särkinen et al. ⁶⁴	2013	BMC Evol Biol	1,076	2,700	39.9	Yes	—	
Cacti	Thompson et al. ⁶⁵	2026	Sci Data	1,063	1,900	55.9	Yes	CoL	
Eucalyptus	Crisp et al. ⁶⁶	2024	J Syst Evol	399	920	43.4	No	—	
Other animals									
Corals	Vaga et al. ³³	2025	Nature	289	1,633	17.7	Yes	WoRMS	
Cephalopods	Basava et al. ⁶⁷	2024	bioRxiv ^e	81	800	10.1	Yes	WoRMS	
Nematodes	Qing et al. ⁶⁸	2025	Syst Biol	166	28,537	0.6	No	CoL	
Bryozoa	Orr et al. ³⁸	2022	Sci Adv	517	6,000	8.6	No	WoRMS	
Bivalves	Pfeiffer et al. ⁴²	2019	MPE	73	20,000	0.4	No	WoRMS	

Group	Study	Year	Journal	Tips	Desc.	%	Dated	Sp. count source
Gastropods	Zapata et al. ⁴³	2014	Proc R Soc B	54	65,000	0.1	No	WoRMS
Sponges	Lavrov et al. ⁶⁹	2023	PLOS ONE	120	9,470	1.3	Yes	WoRMS
Echinoderms	Mongiardino Mongiardino Koch and Thompson ⁷⁰	2022	eLife	66	7,550	0.9	Yes	WoRMS
Microbes & protists								
Diatoms	Nakov et al. ⁷¹	2018	New Phytol	244	20,000	1.2	Yes	AlgaeBase
Bacteria	Parks et al. ²⁹	2025	NAR	136,646	715,230 ^c	19.1	No	GTDB
Archaea	Parks et al. ²⁹	2025	NAR	6,968	17,245 ^c	40.4	No	GTDB
Fungi	Li et al. ³⁹	2021	Curr Biol	1,602	155,000	1.0	No	Sp. Fungorum
Cross-kingdom								
Eukaryote root	Williamson et al. ⁷²	2025	Nature	100	†	†	No ^d	—
TimeTree	Hedges et al. ³⁵	2015	MBE	50,645	†	†	No ^d	—
Tetrapod fam.	Condamine et al. ¹⁴	2019	Ecol Lett	218 trees	†	†	Yes	—
Not yet represented								
Seed plants (strict non-imputed)	—	—	—	0	383,671	0	—	WFO
Birds (strict non-imputed)	—	—	—	0	11,195	0	—	IUCN
Amphibians (strict non-imputed)	—	—	—	0	8,863	0	—	AmphibiaWeb
Squamates (strict non-imputed)	—	—	—	0	11,744	0	—	Reptile DB
Sharks (strict non-imputed)	—	—	—	0	1,282	0	—	Chond. ToL
Acari (mites, ticks)	—	—	—	0	~55,000	0	—	Zhang 2013
Rhodophyta	—	—	—	0	~7,000	0	—	AlgaeBase
Apicomplexa	—	—	—	0	~6,000	0	—	CoL
Myxozoa	—	—	—	0	~2,400	0	—	WoRMS
Rotifera	—	—	—	0	~2,200	0	—	CoL
Euglenozoa	—	—	—	0	~2,000	0	—	CoL
Cercaria	—	—	—	0	~2,000	0	—	CoL
Scorpiones	—	—	—	0	~2,700	0	—	CoL
Tardigrada	—	—	—	0	1,488	0	—	CoL
Acanthocephala	—	—	—	0	~1,300	0	—	CoL
Chrysophyta	—	—	—	0	~1,200	0	—	AlgaeBase
Oomycetes	—	—	—	0	~1,000	0	—	CoL
Other arachnids	—	—	—	0	~12,000	0	—	CoL
Other hexapods	—	—	—	0	~10,000	0	—	CoL
Pycnogonida	—	—	—	0	~1,400	0	—	WoRMS
Other Mollusca	—	—	—	0	~1,500	0	—	WoRMS
Other phyla ⁱ	—	—	—	0	~4,300	0	—	WoRMS, CoL

^aMultiple individuals per species; coverage capped at 100% once the tree spans the recognized species pool. ^bGenus-level. ^cGenome-based species clusters. ^dBranch lengths present but not strictly ultrametric. ^ePreprint. ^fIncludes 21 additional phyla lacking any broadly representative downloadable species-level tree: Gastrotricha (~850 spp.), Xenacoelomorpha (~430), Kinorhyncha (~300), Onychophora (~180), Sipuncula (~130), Chaetognatha (~130), Hemichordata (~130), Dicyemida (~125), Gnathostomulida (~100), Radiolaria (~800), Xanthophyta (~600), Cryptophyta (~200), Loricifera (~40), Phoronida (~20), Priapulida (~25), Orthonectida (~15), Glaucophyta (~15), Placozoa (3), Cycliophora (2), Micrognathozoa (1). [†]Coverage not directly comparable or denominator not available for this dataset; these datasets contribute to the collection but are not used in group-level coverage calculations. Abbreviations: MDD = Mammal Diversity Database; WFO = World Flora Online; CoL = Catalogue of Life; WSC = World Spider Catalog; WoRMS = World Register of Marine Species; MPE = Mol. Phylogenet. Evol.; MBE = Mol. Biol. Evol.; NAR = Nucleic Acids Res. Full per-dataset details (DOIs, download URLs, full provenance) are listed in the supplementary data_provenance.csv and in Supplementary Tables S5 and S7.