Molecular Phylogenetics and Evolution 190 (2024) 107963

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

A supermatrix phylogeny of the world's bees (Hymenoptera: Anthophila)

Patricia Henríquez-Piskulich [a,1,*], Andrew F. Hugall [a,b,1,*], Devi Stuart-Fox [a]

[a] School of BioSciences, The University of Melbourne, Parkville, Victoria, Australia
[b] Department of Sciences, Museums Victoria, Melbourne, Victoria, Australia

ARTICLE INFO

Keywords:
Bees
Molecular phylogenetics
Systematics
Macroevolution

ABSTRACT

The increasing availability of large molecular phylogenies has provided new opportunities to study the evolution of species traits, their origins and diversification, and biogeography; yet there are limited attempts to synthesise existing phylogenetic information for major insect groups. Bees (Hymenoptera: Anthophila) are a large group of insect pollinators that have a worldwide distribution, and a wide variation in ecology, morphology, and life-history traits, including sociality. For these reasons, as well as their major economic importance as pollinators, numerous molecular phylogenetic studies of family and genus-level relationships have been published, providing an opportunity to assemble a bee 'tree-of-life'. We used publicly available genetic sequence data, including phylogenomic data, reconciled to a taxonomic database, to produce a concatenated supermatrix phylogeny for the Anthophila comprising 4,586 bee species, representing 23% of species and 82% of genera. At family, subfamily, and tribe levels, support for expected relationships was robust, but between and within some genera relationships remain uncertain. Within families, sampling of genera ranged from 67 to 100% but species coverage was lower (17–41%). Our phylogeny mostly reproduces the relationships found in recent phylogenomic studies with a few exceptions. We provide a summary of these differences and the current state of molecular data available and its gaps. We discuss the advantages and limitations of this bee supermatrix phylogeny (available online at beetreeoflife.org), which may enable new insights into long standing questions about evolutionary drivers in bees, and potentially insects more generally.

1. Introduction

In the last two decades, multigene supermatrices have been widely used for phylogenetic reconstruction across the diversity of life, opening new opportunities for comparative and macroevolutionary studies of species traits, diversification and biogeography. Within the animal kingdom, these supermatrix phylogenies comprise mainly vertebrate groups such as marsupials, bats, passerine birds, gobies, and cetaceans, to name a few (Amador et al., 2018; Jønsson et al., 2016; McCraney et al., 2020; McGowen et al., 2009; Mitchell et al., 2014). Despite insects constituting most of the species on earth (Stork, 2018), phylogenies with a comprehensive and balanced sampling of known diversity are currently lacking for major insect groups, with the exception of butterflies (Chazot et al., 2021; Foottit and Adler, 2018; Kawahara et al., 2023).

Bees (Hymenoptera: Anthophila) are a large and diverse group of insect pollinators with over twenty thousand described species and a worldwide distribution (Orr et al., 2021). This group has received increased attention in recent decades given their importance as pollinators in the context of insect decline (Potts et al., 2016; Wagner, 2020; Winfree et al., 2009). Bees also hold special interest for understanding life-history evolution (Michener, 2007). Species broadly fall into one of three life-history categories (Danforth et al., 2019): i. Solitary, comprising more than 75% of described species where all females are able to produce offspring and each build and provision their own nest; ii. Social, representing little less than 10%; where species display division of labour, cooperative brood care and generational overlap; iii. Parasitic, including close to 13% of species, where most are brood parasites which lay their eggs in the nest of other bee species, and a few are social parasites where the female replaces the queen, takes over the colony and co-opts worker females to rear her offspring. For some species the lines between these categories are blurred, mainly in the bee family Halictidae, where some species are socially polymorphic (Davison and Field, 2018; McFrederick et al., 2014; Plateaux-Quénu et al., 2000; Richards et al., 2003). Bee species also differ in their morphological features, annual life cycle, diet breadth, and nesting behaviour. The

* Corresponding authors.
E-mail addresses: henriquezpiskulichp@gmail.com (P. Henríquez-Piskulich), ahugall@museum.vic.gov.au (A.F. Hugall).
[1] A.F.H and P.H.-P. contributed equally to this work.

https://doi.org/10.1016/j.ympev.2023.107963
Received 26 June 2023; Received in revised form 28 October 2023; Accepted 4 November 2023
Available online 14 November 2023

latter includes a variety of substrates such as soil, wood, or pithy stems, and the use of diverse materials including their own glandular secretions, and a wide range of foreign resources (e.g., floral oils, leaves, resin, sand, pebbles) (Michener, 2007). The diverse life history and natural history of this remarkable group of insects make them ideal for macroecological and macroevolutionary analyses.

The first large supermatrix phylogeny for bees was by Hedtke et al. (2013), with several more recent attempts (albeit with similar or fewer species: Chesters, 2020; Chesters et al., 2023). Numerous molecular phylogenetic studies of relationships between and/or within families or genera for bees have been published over the years (Supplementary Table S1), increasingly using new and powerful phylogenomic data (e.g., Almeida et al. 2023; Peters et al. 2017). In addition, comprehensive molecular phylogenies have been published for the seven bee families (Melittidae, Andrenidae, Halictidae, Colletidae, Stenotritidae, Megachilidae and Apidae (as defined by Michener (2007)). Although some uncertainties remain, most evolutionary relationships between families and subfamilies are robust (Almeida et al., 2023, 2019, 2012; Bossert et al., 2021, 2019; Danforth et al., 2008; Gonzalez et al., 2012; Litman et al., 2011; Michez et al., 2009; Peters et al., 2011; Sann et al., 2018). Considering the resulting substantial amount of gene and phylogenomic data, it is timely to generate a new time-calibrated bee 'tree-of-life' supermatrix phylogeny.

We used the supermatrix approach (e.g., Driskell et al., 2004) to combine diverse publicly available gene and phylogenomic sequence data, all reconciled to a taxonomic database to ensure nomenclatural consistency. We focused on curating and analysing widely sampled loci used in previously published phylogenies to maximize overlap of data among lineages and reduce supermatrix sparseness. The supermatrix was then used to infer, by maximum likelihood, the largest bee phylogeny to date with 4,586 species (23% of described species and 82% of genera). While this supermatrix phylogeny does not aim to solve current uncertainties in phylogenetic relationships, it provides a single curated synthesis that increases species sampling, which may enable multiple new research opportunities. In addition, it provides information regarding data gaps that need to be addressed to improve resolution for some bee genera. This supermatrix is available for download and can be subsetted through the online tool beetreeoflife.org. In addition, the website also hosts 1,000 bootstrap sample trees converted to dated chronograms to represent the phylogenetic information and uncertainty in the supermatrix.

2. Methods

2.1. Taxonomic database

Large scale phylogenies require a consistent taxonomy and nomenclature to which all data can be reconciled (e.g., Driskell et al., 2004; Hosner et al., 2022; Thomas et al., 2013). For this work, to reconcile binomial nomenclature of the molecular data used to build the supermatrix phylogeny, a taxonomic database was developed for Anthophila based on data provided by Orr et al. (2021), which included synonyms and a more accurate and curated version of current nomenclature than the Open Tree of Life project (Rees and Cranston, 2017). We modified this database to reflect revised generic-level classification from more recent published taxonomic references that have provided generic stability. In addition, final binomial nomenclature was run in the BeeDC R package to flag potential nomenclature issues (Dorey et al., 2023), which were crosschecked against the literature. Our taxonomic database only includes species that are part of the phylogeny, recognises seven bee families and 28 subfamilies, and provides a list of genera, species, and notes on nomenclature decisions and the taxonomic research that supports them (Supplementary file 1). In addition, differences exist regarding nomenclature between some databases (e.g., Integrated Taxonomic Information System, Catálogo Moure para as espécies de abelhas neotropicais, Atlas Hymenoptera, Catalogue of Life), and thus the taxonomic database also includes alternate names for those species with names that have not been updated to follow the most recent nomenclature decisions. Because taxonomy and nomenclature are constantly changing, especially for taxa where relationships are yet to be resolved, the goal of this taxonomic database is to make it easier for researchers to track binomial nomenclature changes in the phylogeny and easily change them if necessary.

2.2. Supermatrix data components

We selected the widely-used supermatrix approach (Chesters, 2020; Driskell et al., 2004; Hedtke et al., 2013) to build a large bee 'tree-of-life' over other composite 'supertree' type methods (e.g., Chesters et al., 2023; Sanderson et al., 1998; Upham et al., 2019). Supermatrix methods use all available genetic sequences to simultaneously estimate relationships, branch length and uncertainty (bootstrap resampling). We built our phylogenetic supermatrix database for bees based on four key components: i. multi-gene sequence data downloaded from global sequence databases (NCBI and BOLD); ii. a phylogenomic dataset to provide a strong backbone across the major lineages, both in topology and relative branch length; iii. published ultra conserved element (UCE) datasets combined and condensed; and iv. nomenclature reconciled to the taxonomic database. Collation and curation of each component was done with custom Bash scripts and Microsoft Excel (see Supplementary file 2 and Supplementary file 3) (Microsoft Corporation, 2023); details for each component are given in the following sections. Given the multiple data sources, bespoke taxonomic database, and the need for hands-on refinement, we expanded our supermatrix assembly system used in previous works (e.g., Hugall and Stuart-Fox, 2012; Oliver et al., 2023).

2.2.1. Multi-gene sequence data

Data were downloaded from the NCBI nucleotide collection (https://www.ncbi.nlm.nih.gov/genbank). We used gene sequences obtained from key phylogenetic references in BLAST+organism searches with an E-value threshold of 5e-6 and retained aligned regions, excluding model and environmental data. For protein coding genes and ribosomal RNA genes we used tblastn and blastn respectively. In total, we obtained data for seven nuclear protein-coding genes (ArgK, CAD, NaK, Pol II, Wnt-1, LW Rh, and EF-1α), two ribosomal genes (16S and 28S), and two mitochondrial protein coding genes (Cytb and COI) (Table 1). We focussed on these widely sampled loci with the greatest proportion of taxa and phylogenetic spread to maximize overlap of data among lineages. This approach makes it easier to check individual genes, and limits the proportion of taxa without data in common, which can distort phylogenetic inference (Freyman, 2015; Sanderson et al., 2010; Wiens and Morrill, 2011).

We focused on protein coding regions (e.g., exons), as they are easier to align, and make it simple to add new data. For the purposes of this work, we did not include the widely sequenced ribosomal 18S gene due to its low and poor phylogenetic signal at this intrafamilial scale (Soltis et al., 1999). Because of the limited coverage of whole mitogenome data for bees (Husemann et al., 2021), we extracted only near full-length mitochondrial genes (16S, Cytb, COI) with the reference BLAST. Rather than selecting one 'best' longest species sequence per locus, we retained all sequence data to first assess the gene tree monophyly of nominate species as well as genera. Due to the great number of sequences, EF-1α, LW Rh and COI genes were split into two sets: Apidae and the rest.

Gene sequences were aligned with MAFFT v. 7.245 (Katoh and Standley, 2013) using the 'AUTO' setting. We then assembled alignments into a custom Microsoft Excel database, and reconciled nomenclature with the taxonomic list (Supplementary file 1). We refined protein coding gene alignments by hand in BioEdit (Hall, 1999) using the original references as a guide, to have all coding regions strictly in-frame. Despite being time consuming, aligning protein coding genes this way makes adding taxa simpler. For the ribosomal RNA genes, we used the MAFFT alignment as is.

We inferred gene trees through IQ-TREE (Nguyen et al., 2015), and in some cases RAxML (Stamatakis, 2014). IQ-TREE used ultrafast

Table 1
Sampling summary across genetic data. Molecular data was sourced from published datasets, NCBI and BOLD.

Gene*	Sites	Genera	Species	p-sp	Description
Nuclear					
Phylogenomic 'stub'	21,546	135	200	0.04	Almeida et al. (2023) UCE phylogenomic alignment subset
UCE 'stub'	13,250	183	678	0.15	Composite UCE data subset (Bossert, Sless, Freitas, and Pisanty)
ArgK	546	62	611	0.13	Arginine kinase gene, exon 2
CAD	450	138	677	0.15	Conserved ATPase domain protein gene exon
EF-1α	1,107	332	2,049	0.45	Elongation factor 1-alpha gene exon
LW Rh	642	318	1,846	0.40	Long wavelength sensitive opsin 1 gene exon
NaK	1,461	233	790	0.17	Sodium potassium adenosine triphosphatase gene exon
Pol II	840	198	853	0.19	RNA polymerase II gene exon
Wnt-1	456	257	841	0.18	Wingless gene exon
28S rDNA	1,440	339	1,253	0.27	Large subunit nuclear ribosomal RNA gene
Mitochondrial					
16S rDNA	522	74	508	0.11	Mitochondrial large subunit ribosomal RNA gene
COI	1,473	330	3,839	0.84	Mitochondrial cytochrome oxidase subunit 1 gene
Cytb	1,047	67	487	0.11	Mitochondrial cytochrome *b* gene
Total	44,780	2,666	14,632		

p-sp: proportion of bee species in the phylogeny with that gene.
Genera with at least one species with gene sequence.
'stub' is an informal description of a reduced subset of originally much larger data.
* For our purposes, 'gene' refers to a discrete genetic data unit.

bootstrap consensus (Hoang et al., 2018) with models of sequence evolution identified by ModelFinder implemented in IQ-TREE (Kalyaanamoorthy et al., 2017). RAxML used fast bootstrap and the GTR+G model, with the standard -f a setting implemented on CIPRES (Miller et al., 2010). We assessed gene trees for non-monophyletic genera and species, and aberrant sequences such as paralogs (e.g., EF-1α and LW Rh variants) and gross misalignments (long terminal branches). This was done via Bash scripts for assessing taxon monophyly, followed by visual inspection of trees and alignments for the final decision. Then, we updated the dataset by removing abnormal sequence data and in some cases revising the nomenclature. Datasets were realigned if necessary.

The mitochondrial COI was processed slightly differently because it is a gene extensively sampled in bees, intra-specific diversity, and because it is often sequenced in separate fragments. We reduced this gene alignment to a single consensus sequence per species, based on the most common base per site (with ties scored as ambiguous). This approach tends towards the most commonly sequenced sub-lineage and is a simple way to combine data, discount rarer aberrant sequences and rationalize choice of intra-specific lineage complexity. The consensus alignment was then subjected to the same procedure of gene tree and genera monophyly assessment as described for the other genes included in this work. Additionally, we included COI data from The Barcode of Life Data System (BOLD, http://www.boldsystems.org) for a selection of species that had multiple samples that clustered closely (>5%), or that represented taxa we already had with other genes if there was only one sequence available, and that fell within the lowest rank (genus or subfamily depending on taxon sampling) in our working COI tree. Information on data sources for all taxa and genes is summarized in Supplementary file 3.

2.2.2. Higher-level phylogenomic backbone

Sparse supermatrices are best if the data is phylogenetically structured across the major lineages but information in global sequence repositories can be quite uneven in this regard (Beaulieu and O'Meara, 2018; Wiens and Morrill, 2011). Fortunately, phylogenomic data is now filling this gap, providing the desired strong backbone to tie together the assemblage of species data. Several suitable family-level datasets have been available for bees (e.g., Branstetter et al., 2017; Peters et al., 2017), but the recent UCE phylogenomic work of Almeida et al. (2023) now provides a broader estimate of the higher level phylogenetic tree for Anthophila, spanning 216 species in all seven families and 28 subfamilies, and most tribes. This length of sequence (830 loci) is unnecessary for our purposes and overly computationally taxing, thus we condensed down the original Bossert and Almeida (2023) from 364.3 kb to a more practical 21.5 kb, by filtering sites by taxon coverage (>74%) then randomly sub-sampling one tenth (jack-knife), thereby reducing the proportion of missing data from the original 79% complete to 87% complete. A phylogeny was then inferred from this phylogenomic 'stub' (as previously described for the gene fragment datasets) and compared to the original published trees (Almeida et al., 2023).

2.2.3. Composite UCE dataset

Currently there are a growing number of published studies of groups of bees using separate UCE datasets that do not overlap in taxa. There are several approaches to combining these into a single common alignment: i. reconstruct from primary reads; ii. reconstruct from locus datasets; iii. reconstruct from the final processed and refined datamatrix. We took the third approach, as a simple compromise that has the benefit of leveraging from previous bioinformatic work, and be preferable to composite 'supertree' approaches (e.g., Chesters et al., 2023; Kimball et al., 2019).

We combined a subset of data from five previously published datamatrices (Bossert, 2021a, 2021b; Freitas et al., 2020; Pisanty et al., 2022; Sless, 2021). First, we produced consensus sequences for each of the five data-matrices (similar approach as described for the COI gene). Then the 1,388 UCE consensus loci obtained from Bossert (2021a) were mapped onto the consensus sequences of the other data-matrices with BLAST v2.13.0, to identify the corresponding orthologous sections. We then selected a subset of suitable UCE loci using several criteria: maximum e-value of 1e-50, >150 sites, and >50% sites matched to the consensus reference. We aligned these loci with MAFFT as previously described for gene fragments, concatenated them into a UCE datamatrix, and reconciled species names with the taxonomic database. This datamatrix (57.9 kb in 94 loci for 777 samples) was then compacted down to an amount adequate to reconstruct key tree shape, without excess computational burden, by filtering sites by taxon coverage (>74%) then further reduced by one third jack-knifing, resulting in a dataset of 13.3 kb; 85% complete. We then produced a tree to compare to the original published UCE phylogenies, and assess for non-monophyletic genera and aberrant sequences as previously described. Details of the phylogenomic data sources are provided in Supplementary file 3.

2.3. Supermatrix assembly and analysis

Once all supermatrix data components were reconciled to the taxonomic database, aligned and assessed for aberrant elements, they were concatenated and used to produce a supermatrix phylogeny. We did this by taking the best (longest accepted) single exemplar sequence per 'gene' per species, for all species with a total of >500 sites of any data.

The supermatrix was then compacted by removing regions with little or no data (<20% taxa with any data per 'gene') and ambiguous alignment regions with Aliscore v2.2 (Misof and Misof, 2009), while maintaining protein coding reading frames.

Sparse, patchy supermatrices often include taxa groups that share no data in common, which can bias phylogenetic inference (Sanderson et al., 2010; Smirnov and Warnow, 2021; Wiens and Morrill, 2011). For example, for blocks of phylogenomic backbone data to work effectively, the taxa in these blocks should also have as many other genes as possible in the rest of the supermatrix. To ameliorate this issue, we implemented minimal taxonomic rank substitution to improve data density on the phylogenomic 'stub', and species with no overlap of data within genera (Pennell et al., 2016). For genera that contained species that did not share any data with any other member, we reassigned the data to the member of the genus with the most data and removed the source species. In total 63 substitutions were made (0.43% of the matrix). Thus, we ensured that all 200 taxa with the phylogenomic stub, and 68% of the 678 species with the UCE stub, also had at least one other gene. Ninety-four species had both phylogenomic and UCE stub data. However, due to our sub-setting approach there was minimal (<2%) duplication of the UCE data (i.e., the same sequence in both phylogenomic and UCE stubs). In effect these 94 species had a combined 21.5 + 13.3 = 34.8 kb of UCE.

We included five outgroups representing four Apoid wasps families, following the classification proposed by Sann et al. (2018) and based on the phylogenomic 'stub' from Almeida et al. (2023) plus a small amount of gene fragments: *Pulverro* (Ammoplanidae), *Bembix* (Bembicidae), *Cerceris* (Philanthidae), *Philanthus* (Philanthidae) and *Tachysphex* (Crabronidae). As gene data were patchy across outgroups, taxonomic rank-substitution was also used to bolster some of these.

To reduce computational burden, the supermatrix was split into three highly supported subsets comprising the families Megachilidae + Andrenidae + Melittidae, Halictidae + Colletidae + Stenotritidae, and Apidae, each sharing a core set of 18 family representatives with maximal data, and the five outgroups. In addition, we extracted one species with the most data for each nominal genus to create a genus-level supermatrix. For each of these supermatrix subsets, phylogenetic inference was done by IQ-TREE with ultrafast bootstrap, using a RAxML starting tree, all nearest-neighbour interchange and refined tree search settings (-allnni, -beps 8, -bnni, -bcor 0.98, -wbtl, -pers 0.4, -nstop 200, –sprrad 8; via CIPRES). IQ-TREE was run for three iterations. The intention was to focus effort on tree support space, in what can be a difficult problem (Shen et al., 2020). The optimal partitioned sequence evolution model was determined with ModelFinder (in IQ-Tree) using the genus-level supermatrix, and this 8-partition model scheme was then used for all remaining analyses (see Supplementary file 3). This ensures a common model irrespective of the amount of data per gene per family subset (i.e., as if all three were analysed simultaneously). The trees from the three family subsets were then joined together again using the common family representatives, for both the consensus tree and the 1,000 bootstrap trees.

2.4. Tree dating calibration

Patchy and missing data can exacerbate unequal/uneven estimates of branch length among lineages (how clock-like the tree appears), over and above inherent rate variation (Roure et al., 2013; Zheng and Wiens, 2015). To reduce this calibration problem, it is necessary to rely on common partition models, and relaxed-clock models. The species-level supermatrices were far too large to properly run Bayesian relaxed-clock methods (Fisher et al., 2022), and the divide and graft approach is complicated and constraining (e.g., Jetz et al., 2012; Upham et al., 2019). Thus, we used the simpler but still widely used Penalised Likelihood Rate Smoothing (PLRS) method (Sanderson, 2002), as implemented in treePL (Smith and O'Meara, 2012). To enforce chronograms to remain fully bifurcating and improve the efficiency of the PLRS iteration, a small length increment was added to any branch of length >0.0005 (amounting to mean 0.43% of the original total tree length, affecting 6.8% of branches). TreePL was run with log_penalty, 'thorough prime' optimized settings and smoothing factor drawn from a uniform distribution 15–30.

There has been considerable variation among published works regarding the crown age for bees (e.g., Almeida et al., 2012; Cardinal et al., 2018; Cardinal and Danforth, 2013; Freitas et al., 2022; Peters et al., 2017; Schwarz et al., 2006). The recent study of Almeida et al. (2023), combining phylogenomic data with Bayesian fossilized birth–death model (FBD) incorporating an extensive fossil dataset, represents the current best overall estimate. Therefore, we calibrated our analysis with a secondary root calibration from Almeida, and a set of 34 node minimum constraints, based on a version of the Cardinal et al. (2018) Table S3 fossil calibration revised with the Almeida et al. (2023) data S1B_fossils information (Supplementary Table S7). We calibrated the bee root to a broad normal distribution of 120 million years ago (mya) SD 6. The normal root calibration and uniform smoothing factor distributions were achieved by randomly drawing values for each of the 1,000 bootstrap replicates. This calibration scheme was used for both the genus-level and the all-species trees. The results were then compared to published ages for a suite of key higher clades.

3. Results and discussion

Our intention was to produce a useful resource, by collating molecular data for bees, summarizing the current state of the data, and providing a serviceable large-scale supermatrix phylogeny. Molecular data are continuously being generated, and in this process higher taxonomy is also being reconciled. As a result, most of the higher ranks are consistent and well supported: family, subfamily, tribe; but to a lesser extent, genera. We kept the tree inference to an efficient approximation, with appropriate computational effort to match the level of precision inherent in the data. The resulting phylogeny synthesises our current understanding of phylogenetic relationships across bee diversity. In the following sections we provide information regarding the current state of the data, the estimated phylogeny, and the uncertainty for the set of phylogenetic trees produced in this work.

The phylogeny includes 4,586 species and 428 genera, which represent 23% of currently described species and 82% of genera, respectively (Fig. 1). In addition, the genera included contain 96% of all described species. This comprises 44.8 kb sites with a total of 21.8 million base characters from 14,651 data elements, median 11% complete (median 3 genes, 2.5 kb site per species). Supplementary file 3 provides complete information on genetic data for the entire supermatrix. We took a more hands-on approach to curating suitable data rather than an all-in automation to build our dataset (Beaulieu and O'Meara, 2018), with several iterations of collation and summary inference guiding its development. Our dataset is not exhaustive but a substantial representation of the total possible molecular data, as of mid 2023, providing the largest supermatrix dataset and tree of bees to date (Chesters, 2020; Chesters et al., 2023; Hedtke et al., 2013). Tables 1, 2 and Supplementary Table S2 show brief summaries of our sampling by family and by gene. A more detailed breakdown by subfamily is provided in Supplementary Table S3. Within families, genera sampling was high (73 to 100%) but much lower for species (19–41%) (Table 2). Taxonomic sampling is key for phylogenetic inference accuracy (Nabhan and Sarkar, 2012), and while family, subfamily and tribe level relationships are robust, further work is necessary to fill in the gaps and reduce uncertainties in relationships between and within genera.

3.1. Geographic distribution of sampling

Species in the tree are biased towards the Neartic, followed by the Paleartic and Neotropics (Fig. 2, Supplementary Table S4), indicating knowledge gaps in bee distribution and under-sampling of all other regions (see Supplementary file 4 for a description of the methods used to obtain bee distribution data). Despite bees being a diverse insect group

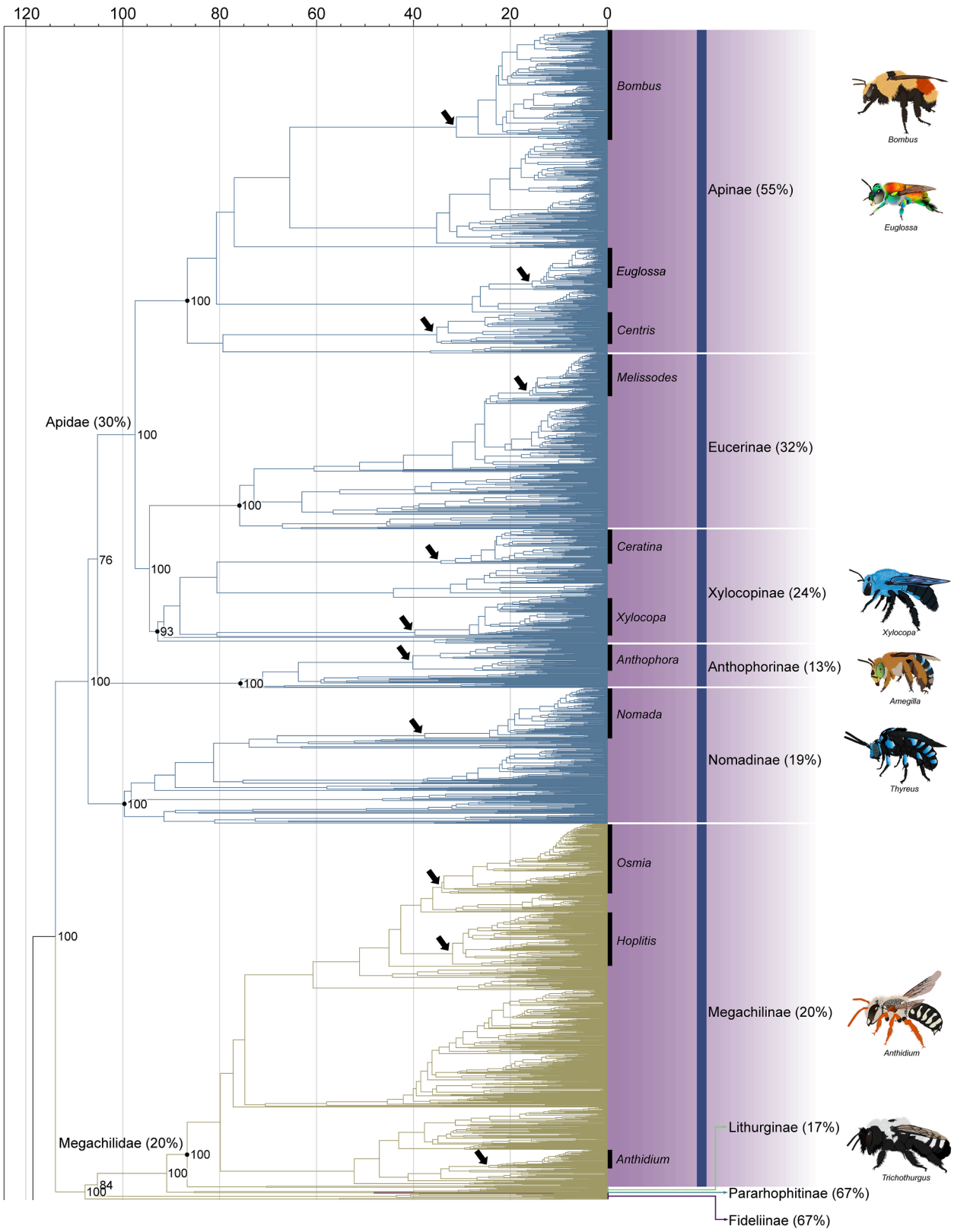

Fig. 1. Supermatrix phylogeny for Anthophila built with public and published sequence data. The full tree features 4,586 species in 428 genera, representing 23% of species and 82% of genera, inferred with IQ-TREE. Root age is based on the work of Almeida et al. (2023) and dating was done with treePL (see Table S7 for calibrations). Taxon coverage is denoted in brackets, and node labels represent ultrafast bootstrap values as a percentage. For more information see the genus level tree in Supplementary Fig. S1. Dots on nodes demark the subfamily clades, and arrows the named genera. Some genera highlighted include bumblebees (*Bombus*) which are one of the best-studied bee taxa in the world; orchid bees in part (*Euglossa*), for their colourful appearance; and two of the largest genera of bees, the sweat bees (*Lasioglossum*) and mining bees (*Andrena*). Illustrations to the right of the tree show some of the genera included in the phylogeny. Family colour coding used throughout. This tree can be downloaded at beetreeoflife.org, which also hosts 1,000 bootstrap sample trees converted to dated chronograms to represent the phylogenetic information and uncertainty in our supermatrix.

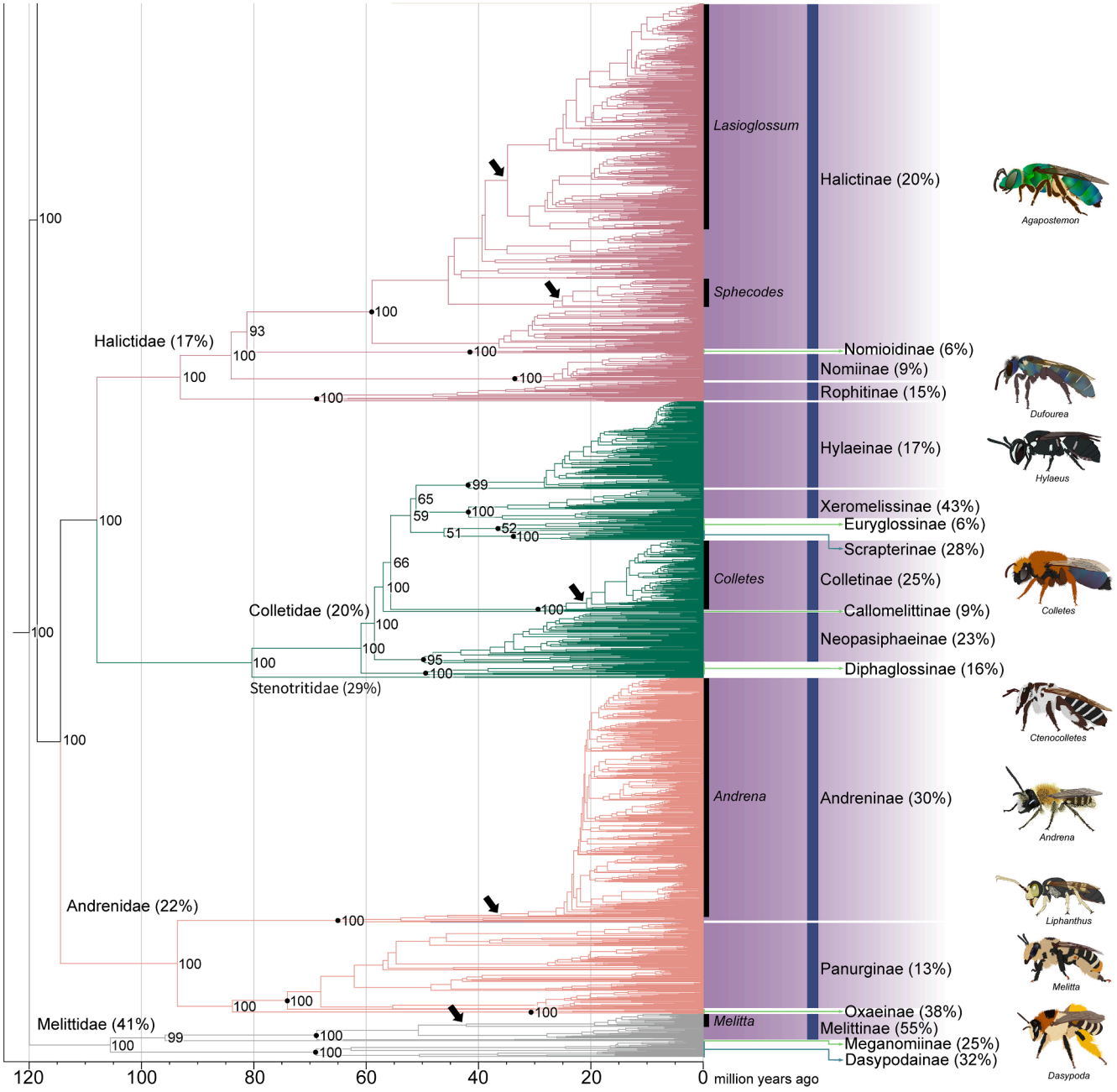

Fig. 1. (*continued*).

with global distribution (Danforth et al., 2019; Michener, 2007), geographical biases remain. This has previously been discussed in terms of ecological gaps (Archer et al., 2014), and taxonomic gaps (Orr et al., 2021), but efforts should also be made to obtain molecular data from other regions to better understand the relationships and the evolution of bees, especially considering that some of these less studied regions might host high bee species richness (Batley and Hogendoorn, 2009; Eardley et al., 2009; Freitas et al., 2009; Groom and Schwarz, 2011; Melin and Colville, 2019).

3.2. Gene sampling

Overall, the data were highly skewed with many species (36%) having one gene (86% COI, 13% UCE stub), and only a small number (16%) having more than five (Fig. 3A-B). However, data were well distributed across lineages and a small number of species with many genes were phylogenetically spread across the tree (Fig. 3A-C, Supplementary Fig. S3, Supplementary Table S2). In addition, the use of shortened subsets of phylogenomic data provided sufficient information to ensure the tree was consistent with published higher-level phylogenies in topology and relative branch length, mostly matching previously published results (see Section 3.4. and Supplementary file 5). Thus, the addition of thousands of taxa with little data to the core backbone had limited effect on the underlying backbone of the tree. This can be seen in the similarity of the genus representative tree and the equivalent subtree pruned from the full all-species tree where 92% of nodes were recovered; remaining nodes had low support in either or both trees (Supplementary Fig. S1). Branch lengths were also very similar, as are the inferred clade ages (see section below).

In total, median IQTree ultrafast bootstrap support value (ufbs) was 86%, with 30% of nodes below 70%. This varied across the family subtrees (Fig. 3D). Apidae and Melittidae had the best support (median

Table 2
Summary of taxonomic sampling by family.

Family	Total subfamilies*	Total genera	Included genera (%)	Total species	Included species (%)
Andrenidae	3	50	44 (88)	2,957	639 (22)
Apidae	5	211	184 (87)	5,829	1,751 (30)
Colletidae	8	76	66 (87)	2,616	520 (20)
Halictidae	4	82	61 (74)	4,400	758 (17)
Megachilidae	4	83	61 (73)	4,099	830 (20)
Melittidae	3	15	10 (67)	201	82 (41)
Stenotritidae	1	2	2 (100)	21	6 (29)
Total	28	519	428 (82)	20,123	4,586 (23)

* All subfamilies are included in this work and total numbers for each rank are based on our taxonomic database.

ufbs 91/93%), and Colletidae the least (median 79%). The genus-level representative tree returned higher support: median ufbs 99% (Fig. 3D, Supplementary Fig. S1).

3.3. Data structure limitations

A key issue in supermatrix data structure is how much data are shared between true sister taxa, a simple measure being the genes taxa have in common (hereafter referred to as 'GIC' – genes in common). GIC is very similar to taxon triplets measures (Sanderson et al., 2010), and provides an indication of data structure limitations by identifying groups that are most affected by scarce data overlap, and would therefore profit from targeted data gap filling (see Freyman, 2015). In a cladistic analysis, taxa with no data in common cannot be recovered as sister taxa (Sanderson et al., 2010). In principle, a supermatrix would have at least one GIC to all species (typically mtDNA) with the remaining genes (typically nuclear) limited to key representatives of the major lineages. However, in practice supermatrices tend to be patchy with much missing data, potentially leaving many true sister taxa with little or no data in common. Nonetheless, the most common gene COI covers 3,892 species (84%, Table 1). The nuclear genes EF-1a and LW Rh are also widely sampled (44 and 40%, respectively).

At family and subfamily levels, data were highly robust and structured with a minimum of seven GIC (median of nine) among all subfamilies (Supplementary Table S5, Fig. 3C). Within genera the overlap of data varied (Supplementary Table S5, Supplementary Fig. S2A-B), with numerous cases of zero shared data between individuals within a genus (Supplementary Table S6; gic0 = proportion of pairwise comparisons with zero GIC). A few examples illustrate some issues: The Andrenidae genera have robust backbones from UCE data but the problem is linking this to taxa with the standard 'legacy' genes. In the large genus *Megachile* (207 species) it is unclear if and where the moderate gic0 (0.12) might affect the tree but in *Stelis* it must be substantial despite the apparently high bootstrap support (Sanderson et al., 2015). The generally increasing GIC deeper into the tree and variation in GIC towards the tips is apparent from GIC mapped onto the tree (Supplementary Fig. S2C; note that in the tree GIC should be >0). Comparing a few groups, *Lassioglossum* and *Nomada* have generally few genes in common (1) but this is quite consistent among taxa hence low proportion of gic0 (species/gic0; *Lassioglossum*: 430/0.003; *Nomada*: 111/0.05; see Supplementary Table S6 for details). On the other hand, *Hylaeus* and *Calliopsis* are poorly structured with a significant minority of gic0 (154/0.18 and 17/0.24 respectively). There are two approaches when dealing with limited GIC, either remove species or fill in the existing gaps. Excluding rogue taxa may also help but this can be a complicated process (Smith, 2022). Until these gaps are filled, species relationships within poorly structured genera should be treated with caution, and in the cases where analyses are sensitive to tree topology, these genera could be pruned back to single genus representatives.

3.4. Similarity to previous studies

As intended, our tree largely reproduces the recent best phylogenomic scale results for families and subfamilies (e.g., Bossert et al., 2021,

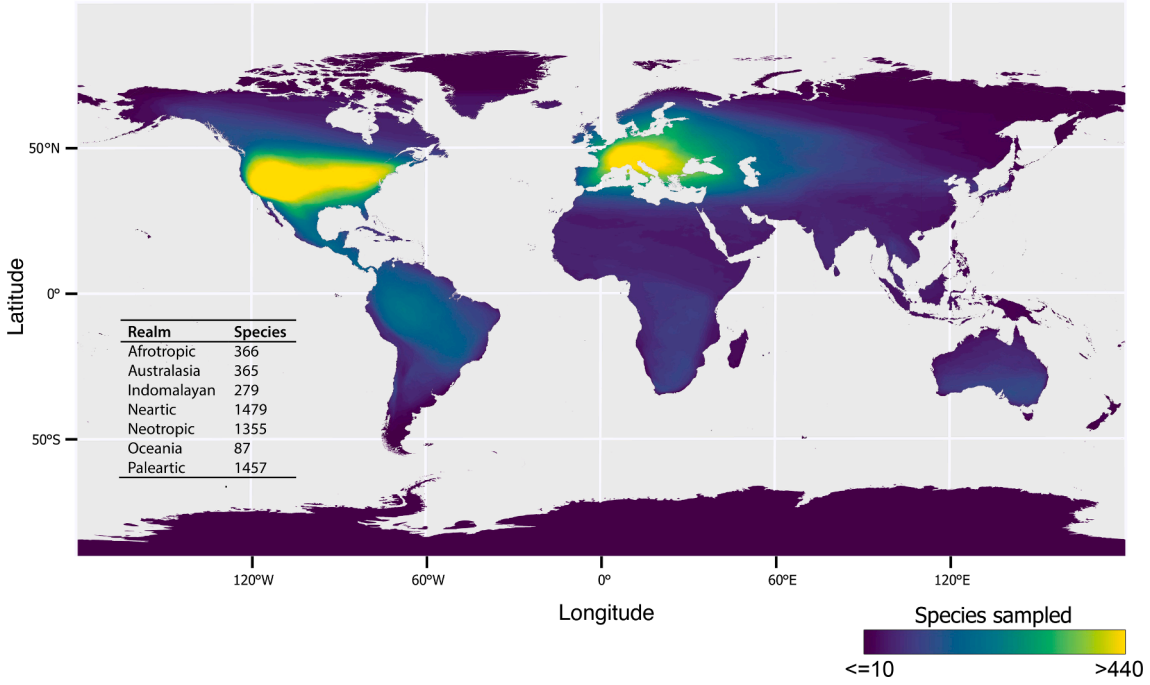

Fig. 2. Distribution of bees with available molecular data that was used to build the supermatrix phylogeny.

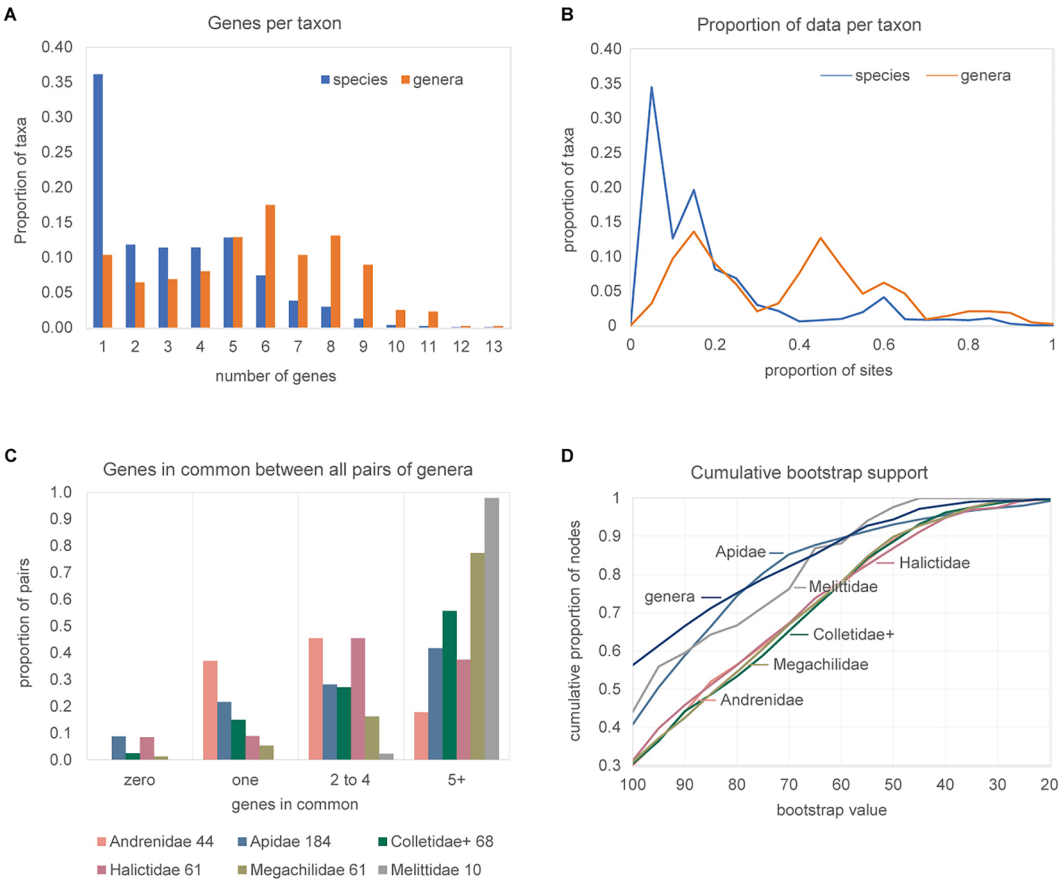

Fig. 3. Graphical summaries of data completeness and tree support. **A.** Number of genetic data elements ('genes') per species in the 4,586 species tree and per genus in the 428 genus representative tree. **B.** Proportion of sites with data in the supermatrix not including the phylogenomic stub. **C.** Number of genes in common between all genera within each family, summarized in four categories (number of genera after name, and in histogram left to right in alphabetical order). **D.** Cumulative bootstrap support across nodes by family and the genus representative tree. Colletidae+ includes Stenotritidae.

2019; Branstetter et al., 2017; Peters et al., 2017; Sann et al., 2018; Sless et al., 2022). These differ somewhat to older multi-gene studies, the data of which we also included (Almeida et al., 2012; Cardinal et al., 2018; Cardinal and Danforth, 2013; Ramos et al., 2022; Rehan et al., 2010). All family-level relationships had 100% ufbs support. Regarding higher groups with more than one taxon in our tree: 25 out of 26 subfamilies were monophyletic, all but Euryglossinae with >95% ufbs support. Fideliinae appeared paraphyletic, as also seen in Almeida et a. (2023); 58 out of 60 tribes were monophyletic (with >90% ufbs); the exceptions, Centridini and Ancyloscelidini, had equivocal support (monophyletic in the species tree, paraphyletic in the genus tree) with weak bootstrap values (see Supplementary file 5 for details), in common with previous phylogenomic studies (Supplementary Table S1). Of 264 genera with more than one species in our tree, nominally 63 were not monophyletic. These non-monophyletic genera fall into four broad categories noted here and discussed further below: i. not monophyletic (Supplementary file 5; but this can depend on the chosen nomenclature); ii. limited resolution (i.e., they are recovered in a minority of bootstraps); iii. artefacts of data structure (but most do not appear to have that problem – see Supplementary Table S5 and S6); iv. artefacts of imperfect ML tree search inference (ameliorated by successive development of the data-matrix culling some highly unstable 'rogue' taxa).

The differences between our full species tree and our genus representative tree mirror the uncertainties reported in some of the primary studies. Most of these differences involve short internodes between lineages. Even with massive amounts of multi-locus UCE data such situations can be unresolvable and/or uninformative (Degnan and Rosenberg, 2009; Hahn and Nakhleh, 2016). Some key remaining uncertainties include the subgeneric relationships within *Andrena* (Andrenidae: Andreninae) given its high degree of paraphyly and polyphyly, where classification changes will likely occur with further efforts to resolve such a diverse and complex genus (Pisanty et al., 2022). A similar situation applies to the enormous genus *Lasioglossum* (Halictidae: Halictinae). The position of the subfamily Anthophorinae (Apidae) is also uncertain, having been recovered as sister to Nomadinae (Bossert et al., 2019), as well as sister to all remaining Apidae subfamilies (Orr et al., 2022). We get yet a third result with the Nomadinae as sister to the Anthophorinae+all other Apidae in both our species and genus trees; but again with short internodes and equivocal support (ufbs 76/92%). We summarise key differences between our results and previously published work in Supplementary file 5, which also mentions some genera that are potentially not monophyletic.

3.5. Dating results

For large phylogenies, Bayesian dating methods are computationally intensive. Penalised likelihood rate smoothing (PLRS), as implemented in TreePL, is an efficient alternative without the complexities of the divide and graft approach (e.g., Jetz et al., 2012; Smirnov and Warnow, 2021; Upham et al., 2019), albeit with more limited scope for incorporating calibration information compared to Bayesian methods. We compared a range of groups across families, for which there were multiple previous estimates from 21 different published studies (Supplementary Fig. S3, Supplementary Table S8). There is considerable variation between studies but in the context of limited precision inherent in relaxed-clock dating (Mello and Schrago, 2014; Sanderson, 2002; Yang and Rannala, 2006), our dating is consistent with the emerging broad overall consensus. In particular, our ages tend to be

slightly older (~11%) than Almeida et al. (2023) but within the confidence interval ranges (Fig. S1B; Bossert and Almeida, 2023).

3.6. Direction of future efforts

Considering the amount of UCE data that has become available in recent years, we anticipate large projects that will consolidate all primary UCE. With these data it would be possible to build a more comprehensive bee backbone 'stub'. In the case of data structure commonality, a targeted approach would be the best way to overcome data gaps, such as the one recently used by Kawahara et al. (2023) for butterflies. Gaps in bee molecular data that emerged from this work are outlined in our summaries of data patchiness. At present most genera are represented, but within genera, species sampling is uneven, ranging between zero to 100% but often below 20%. Proportional molecular data representation is necessary to recover phylogenetic relationships accurately (Dell'Ampio et al., 2014). Thus, to resolve uncertainty in relationships between and within bee genera, future work would benefit from sampling genera that are currently underrepresented. A few key 'legacy genes' would be needed, at least initially, such as COI and EF-1α. Substantial mtDNA often comes with raw UCE data ('genome skimming'; e.g., Branstetter et al. (2021), Sann et al. (2021)), so it is worthwhile including this in order to link species with UCE to the large COI barcode database.

It is important to recognise the limitations of this phylogeny. For most bee species, there is no molecular data available and thus, their placement and relationships remain uncertain. Studies focusing on patterns of diversification in space and time have overcome incomplete sampling by using the birth–death polytomy resolvers of Kuhn et al. (2011) and Thomas et al. (2013), which assign a position to species with no molecular data into phylogenetic trees using information based on taxonomy. This approach has been used to simulate the phylogenetic positions of taxa in mammals and birds (Jetz et al., 2012; Rabosky et al., 2018; Upham et al., 2019), and while the resulting phylogenies might be reliable for diversification and phylogenetic distinctiveness analyses (but see Chang et al., 2020), they should be used with caution in phylogenetic comparative analyses because the placement of these taxa within a phylogeny does not take into account their species trait values (Rabosky, 2015). Our supermatrix only includes species with molecular data, accounting for 23% of the total number of bee species currently accepted. This is far from the coverage of molecular data available for birds or marine fishes (Jetz et al., 2012; Rabosky et al., 2018), but close to the proportion of seed plants with molecular data (Smith and Brown, 2018). For comparative analyses this supermatrix phylogeny provides enough species for multiple phylogenetically independent comparisons (Supplementary Table S4). Nonetheless, we stress the need to use approaches that deal with missing data (Garamszegi and Møller, 2011), and the importance of always assessing the assumptions and biases of phylogenetic comparative analyses, to avoid poor model fits and misinterpretation of results (Cooper et al., 2016; Rangel et al., 2015). Future efforts should focus on sampling species from realms, mainly from the Global South, that currently have a limited amount of molecular data available.

4. Conclusions

This work presents the most species comprehensive bee phylogeny to date, providing a summary of substantial existing sequence data available up to mid 2023, as well as its gaps. The supermatrix phylogeny was built with carefully curated public data, yielding a composite of previously published phylogenetic hypotheses regarding bees that can be downloaded and subsetted online at beetreeoflife.org. Phylogenetic trees in this work show sensible support between bee families, most subfamilies and tribes. At the genus level, genera are for the most part represented, but within genera low data coverage remains a problem in some taxa. Future additional work could improve resolution, particularly for genera that remain under-sampled. This in turn could provide a better understanding of the evolution of this diverse group of insects.

CRediT authorship contribution statement

Patricia Henríquez-Piskulich: Conceptualization, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Andrew F. Hugall:** Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Devi Stuart-Fox:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data and supplementary files generated for this study have been deposited at the Dryad Digital Repository (https://doi.org/10.5061/dryad.80gb5mkw1).

Acknowledgements

We thank the anonymous reviewers for their constructive comments, which greatly improved this manuscript. Patricia Henríquez-Piskulich was supported by funding awarded from the Agencia Nacional de Investigación y Desarrollo de Chile (Scholarship ID 72210037). Devi Stuart-Fox was funded by an Australian Research Council (ARC) Future Fellowship (FT180100216). We thank James Dorey for providing feedback regarding the binomial nomenclature used in this work and running our taxonomic database in the BeeDC package.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2023.107963.

References

Almeida, E.A.B., Pie, M.R., Brady, S.G., Danforth, B.N., 2012. Biogeography and diversification of colletid bees (Hymenoptera: Colletidae): Emerging patterns from the southern end of the world. J. Biogeogr. 39, 526–544. https://doi.org/10.1111/j.1365-2699.2011.02624.x.

Almeida, E.A.B., Packer, L., Melo, G.A.R., Danforth, B.N., Cardinal, S.C., Quinteiro, F.B., Pie, M.R., 2019. The diversification of neopasiphaeine bees during the Cenozoic (Hymenoptera: Colletidae). Zool. Scr. 48, 226–242. https://doi.org/10.1111/zsc.12333.

Almeida, E.A.B., Bossert, S., Danforth, B.N., Porto, D.S., Freitas, F.V., Davis, C.C., Murray, E.A., Blaimer, B.B., Spasojevic, T., Ströher, P.R., Orr, M.C., Packer, L., Brady, S.G., Kuhlmann, M., Branstetter, M.G., Pie, M.R., 2023. The evolutionary history of bees in time and space. Curr. Biol. 33, 3409–3422.e6. https://doi.org/10.1016/j.cub.2023.07.005.

Amador, L.I., Moyers Arévalo, R.L., Almeida, F.C., Catalano, S.A., Giannini, N.P., 2018. Bat systematics in the light of unconstrained analyses of a comprehensive molecular supermatrix. J. Mamm. Evol. 25, 37–70. https://doi.org/10.1007/s10914-016-9363-8.

Archer, C.R., Pirk, C.W.W., Carvalheiro, L.G., Nicolson, S.W., 2014. Economic and ecological implications of geographic bias in pollinator ecology in the light of pollinator declines. Oikos 123, 401–407. https://doi.org/10.1111/j.1600-0706.2013.00949.x.

Batley, M., Hogendoorn, K., 2009. Diversity and conservation status of native Australian bees. Apidologie 40, 347–354. https://doi.org/10.1051/apido/2009018.

Beaulieu, J.M., O'Meara, B.C., 2018. Can we build it? Yes we can, but should we use it? Assessing the quality and value of a very large phylogeny of campanulid angiosperms. Am. J. Bot. 105, 417–432. https://doi.org/10.1002/ajb2.1020.

[dataset] Bossert, S., 2021a. Phylogeny, biogeography, and diversification of the mining bee family Andrenidae. figshare, v3. https://doi.org/10.6084/m9.figshare.14593629.v3.

[dataset] Bossert, S., 2021b. Gene tree estimation error with ultraconserved elements: An empirical study on Pseudapis bees. Dryad. https://doi.org/10.5061/dryad.z08kprrb6.

[dataset] Almeida, E., Bossert, S., 2023. The Evolutionary History of Bees in Time and Space. Mendeley Data, v1. https://doi.org/10.17632/j233njx65x.1.

Bossert, S., Murray, E.A., Almeida, E.A.B., Brady, S.G., Blaimer, B.B., Danforth, B.N., 2019. Combining transcriptomes and ultraconserved elements to illuminate the

phylogeny of Apidae. Mol. Phylogenet. Evol. 130, 121–131. https://doi.org/10.1016/j.ympev.2018.10.012.
Bossert, S., Wood, T.J., Patiny, S., Michez, D., Almeida, E.A.B., Minckley, R.L., Packer, L., Neff, J.L., Copeland, R.S., Straka, J., Pauly, A., Griswold, T., Brady, S.G., Danforth, B. N., Murray, E.A., 2021. Phylogeny, biogeography and diversification of the mining bee family Andrenidae. Syst. Entomol. 47, 283–302. https://doi.org/10.1111/syen.12530.
Branstetter, M.G., Danforth, B.N., Pitts, J.P., Faircloth, B.C., Ward, P.S., Buffington, M.L., Gates, M.W., Kula, R.R., Brady, S.G., 2017. Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees. Curr. Biol. 27, 1019–1025. https://doi.org/10.1016/j.cub.2017.03.027.
Branstetter, M.G., Müller, A., Griswold, T.L., Orr, M.C., Zhu, C.D., 2021. Ultraconserved element phylogenomics and biogeography of the agriculturally important mason bee subgenus *Osmia* (*Osmia*). Syst. Entomol. 46, 453–472. https://doi.org/10.1111/syen.12470.
Cardinal, S., Buchmann, S.L., Russell, A.L., 2018. The evolution of floral sonication, a pollen foraging behavior used by bees (Anthophila). Evolution 72, 590–600. https://doi.org/10.1111/evo.13446.
Cardinal, S., Danforth, B.N., 2013. Bees diversified in the age of eudicots. Proc. Royal Soc. B 280, 20122686. https://doi.org/10.1098/rspb.2012.2686.
Chang, J., Rabosky, D.L., Alfaro, M.E., 2020. Estimating diversification rates on incompletely sampled phylogenies: Theoretical concerns and practicalsolutions. Syst. Biol. 69, 602–611. https://doi.org/10.1093/sysbio/syz081.
Chazot, N., Condamine, F.L., Dudas, G., Peña, C., Kodandaramaiah, U., Matos-Maraví, P., Aduse-Poku, K., Elias, M., Warren, A.D., Lohman, D.J., Penz, C.M., DeVries, P., Fric, Z.F., Nylin, S., Müller, C., Kawahara, A.Y., Silva-Brandão, K.L., Lamas, G., Kleckova, I., Zubek, A., Ortiz-Acevedo, E., Vila, R., Vane-Wright, R.I., Mullen, S.P., Jiggins, C.D., Wheat, C.W., Freitas, A.V.L., Wahlberg, N., 2021. Conserved ancestral tropical niche but different continental histories explain the latitudinal diversity gradient in brush-footed butterflies. Nat. Commun. 12, 5717. https://doi.org/10.1038/s41467-021-25906-8.
Chesters, D., 2020. The phylogeny of insects in the data-driven era. Syst. Entomol. 45, 540–551. https://doi.org/10.1111/syen.12414.
Chesters, D., Ferrari, R.R., Lin, X., Orr, M.C., Staab, M., Zhu, C.D., 2023. Launching insectphylo.org; a new hub facilitating construction and use of synthesis molecular phylogenies of insects. Mol. Ecol. Resour. 23, 1556–1573. https://doi.org/10.1111/1755-0998.13817.
Cooper, N., Thomas, G.H., FitzJohn, R.G., 2016. Shedding light on the 'dark side' of phylogenetic comparative methods. Methods Ecol. Evol. 7, 693–699. https://doi.org/10.1111/2041-210X.12533.
Danforth, B.N., Eardley, C., Packer, L., Walker, K., Pauly, A., Randrianambinintsoa, F.J., 2008. Phylogeny of halictidae with an emphasis on endemic african halictinae. Apidologie 39, 86–101. https://doi.org/10.1051/apido:2008002.
Danforth, B., Minckley, R., Neff, J., 2019. The Solitary Bees: Biology, Evolution, Conservation. Princeton University Press, first ed., Princeton, New Jersey.
Davison, P.J., Field, J., 2018. Limited social plasticity in the socially polymorphic sweat bee *Lasioglossum calceatum*. Behav. Ecol. Sociobiol. 72, 56. https://doi.org/10.1007/s00265-018-2475-9.
Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24, 332–340. https://doi.org/10.1016/j.tree.2009.01.009.
Dell'Ampio, E., Meusemann, K., Szucsich, N.U., Peters, R.S., Meyer, B., Borner, J., Petersen, M., Aberer, A.J., Stamatakis, A., Walzl, M.G., Minh, B.Q., Von Haeseler, A., Ebersberger, I., Pass, G., Misof, B., 2014. Decisive data sets in phylogenomics: Lessons from studies on the phylogenetic relationships of primarily wingless insects. Mol. Biol. Evol. 31, 239–249. https://doi.org/10.1093/molbev/mst196.
Dorey, J.B., Fischer, E.E., Chesshire, P.R., Nava-Bolaños, A., O'Reilly, R.L., Bossert, S., Collins, S.M., Lichtenberg, E.M., Tucker, E.M., Smith-Pardo, A., Falcon-Brindis, A., Guevara, D.A., Ribeiro, B., Pedro, D. de, Pickering, J., Hung, K.-L.J., Parys, K.A., McCabe, L.M., Rogan, M.S., Minckley, R.L., Velazco, S.J.E., Griswold, T., Zarrillo, T. A., Jetz, W., Sica, Y. V, Orr, M.C., Guzman, L.M., Ascher, J.A., Hughes, A.C., Cobb, N. S., 2023. BeeDC: An R package and globally synthesised and flagged bee occurrence dataset, bioRxiv 2023.06.30.547152. https://doi.org/10.1101/2023.06.30.547152.
Driskell, A.C., Ané, C., Burleigh, J.G., McMahon, M.M., O'Meara, B.C., Sanderson, M.J., 2004. Prospects for building the tree of life from large sequence databases. Science 306, 1172–1174. https://doi.org/10.1126/science.1102036.
Eardley, C.D., Gikungu, M., Schwarz, M.P., 2009. Bee conservation in Sub-Saharan Africa and Madagascar: Diversity, status and threats. Apidologie 40, 355–366. https://doi.org/10.1051/apido/2009016.
Fisher, A.A., Hassler, G.W., Ji, X., Baele, G., Suchard, M.A., Lemey, P., 2022. Scalable Bayesian phylogenetics. Philos. Trans. Royal. Soc. B 377, 20210242. https://doi.org/10.1098/rstb.2021.0242.
[dataset] Freitas, F., Branstetter, M.G., Griswold, T., Almeida, E.A.B., 2020. Partitioned gene-tree analyses and gene-based topology testing help resolve incongruence in a phylogenomic study of host-specialist bees (Apidae: Eucerinae). Zenodo, v2. https://doi.org/10.5281/zenodo.3996596.
Foottit, R.G., Adler, P.H., 2018. In: Insect biodiversity: science and society, first ed. 2. John Wiley & Sons, first ed. Hoboken, New Jersey.
Freitas, F.V., Branstetter, M.G., Casali, D.M., Aguiar, A.J.C., Griswold, T., Almeida, E.A. B., 2022. Phylogenomic dating and Bayesian biogeography illuminate an antitropical pattern for eucerine bees. J. Biogeogr. 49, 1034–1047. https://doi.org/10.1111/jbi.14359.
Freitas, B.M., Imperatriz-Fonseca, V.L., Medina, L.M., Kleinert, A.D.M.P., Galetto, L., Nates-Parra, G., Javier, J., 2009. Diversity, threats and conservation of native bees in the Neotropics. Apidologie 40, 332–346. https://doi.org/10.1051/apido/2009012.
Freyman, W.A., 2015. SUMAC: Constructing phylogenetic supermatrices and assessing partially decisive taxon coverage. Evol. Bioinforma. 11, 263–266. https://doi.org/10.4137/EBO.S35384.
Garamszegi, L.Z., Møller, A.P., 2011. Nonrandom variation in within-species sample size and missing data in phylogenetic comparative studies. Syst. Biol. 60, 876–880. https://doi.org/10.1093/sysbio/syr060.
Gonzalez, V.H., Griswold, T., Praz, C.J., Danforth, B.N., 2012. Phylogeny of the bee family Megachilidae (Hymenoptera: Apoidea) based on adult morphology. Syst. Entomol. 37, 261–286. https://doi.org/10.1111/j.1365-3113.2012.00620.x.
Groom, S.V.C., Schwarz, M.P., 2011. Bees in the southwest pacific: Origins, diversity and conservation. Apidologie 42, 759–770. https://doi.org/10.1007/s13592-011-0079-8.
Hahn, M.W., Nakhleh, L., 2016. Irrational exuberance for resolved species trees. Evolution 70, 7–17. https://doi.org/10.1111/evo.12832.
Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41, 95–98.
Hedtke, S.M., Patiny, S., Danforth, B.N., 2013. The bee tree of life: A supermatrix approach to apoid phylogeny and biogeography. BMC Evol. Biol. 13, 138. https://doi.org/10.1186/1471-2148-13-138.
Hoang, D.T., Chernomor, O., Von Haeseler, A., Minh, B.Q., Vinh, L.S., 2018. UFBoot2: Improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35, 518–522. https://doi.org/10.1093/molbev/msx281.
Hosner, P.A., Zhao, M., Kimball, R.T., Braun, E.L., Burleigh, J.G., 2022. Updating splits, lumps, and shuffles: Reconciling GenBank names with standardized avian taxonomies. Ornithology 139, ukac045. https://doi.org/10.1093/ornithology/ukac045.
Hugall, A.F., Stuart-Fox, D., 2012. Accelerated speciation in colour-polymorphic birds. Nature 485, 631–634. https://doi.org/10.1038/nature11050.
Husemann, M., Neiber, M.T., Nickel, J., Reinbold, C.V.M., Kuhlmann, M., Cordellier, M., 2021. Mitogenomic phylogeny of bee families confirms the basal position and monophyly of Melittidae. Zool. Scr. 50, 352–357. https://doi.org/10.1111/zsc.12468.
Jetz, W., Thomas, G.H., Joy, J.B., Hartmann, K., Mooers, A.O., 2012. The global diversity of birds in space and time. Nature 491, 444–448. https://doi.org/10.1038/nature11631.
Jønsson, K.A., Fabre, P.H., Kennedy, J.D., Holt, B.G., Borregaard, M.K., Rahbek, C., Fjeldså, J., 2016. A supermatrix phylogeny of corvoid passerine birds (Aves: Corvides). Mol. Phylogenet. Evol. 94, 87–94. https://doi.org/10.1016/j.ympev.2015.08.020.
Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., Jermiin, L.S., 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14, 587–589. https://doi.org/10.1038/nmeth.4285.
Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30, 772–780. https://doi.org/10.1093/molbev/mst010.
Kawahara, A.Y., Storer, C., Carvalho, A.P.S., Plotkin, D.M., Condamine, F.L., Braga, M.P., Ellis, E.A., St Laurent, R.A., Li, X., Barve, V., Cai, L., Earl, C., Frandsen, P.B., Owens, H.L., Valencia-Montoya, W.A., Aduse-Poku, K., Toussaint, E.F.A., Dexter, K. M., Doleck, T., Markee, A., Messcher, R., Nguyen, Y.-L., Badon, J.A.T., Benítez, H.A., Braby, M.F., Buenavente, P.A.C., Chan, W.-P., Collins, S.C., Rabideau Childers, R.A., Dankowicz, E., Eastwood, R., Fric, Z.F., Gott, R.J., Hall, J.P.W., Hallwachs, W., Hardy, N.B., Sipe, R.L.H., Heath, A., Hinolan, J.D., Homziak, N.T., Hsu, Y.-F., Inayoshi, Y., Itliong, M.G.A., Janzen, D.H., Kitching, I.J., Kunte, K., Lamas, G., Landis, M.J., Larsen, E.A., Larsen, T.B., Leong, J.V., Lukhtanov, V., Maier, C.A., Martinez, J.I., Martins, D.J., Maruyama, K., Maunsell, S.C., Mega, N.O., Monastyrskii, A., Morais, A.B.B., Müller, C.J., Naive, M.A.K., Nielsen, G., Padrón, P. S., Peggie, D., Romanowski, H.P., Sáfián, S., Saito, M., Schröder, S., Shirey, V., Soltis, D., Soltis, P., Sourakov, A., Talavera, G., Vila, R., Vlasanek, P., Wang, H., Warren, A.D., Willmott, K.R., Yago, M., Jetz, W., Jarzyna, M.A., Breinholt, J.W., Espeland, M., Ries, L., Guralnick, R.P., Pierce, N.E., Lohman, D.J., 2023. A global phylogeny of butterflies reveals their evolutionary history, ancestral hosts and biogeographic origins. Nat. Ecol. Evol. 7, 903–913. https://doi.org/10.1038/s41559-023-02041-9.
Kimball, R.T., Oliveros, C.H., Wang, N., White, N.D., Barker, F.K., Field, D.J., Ksepka, D. T., Chesser, R.T., Moyle, R.G., Braun, M.J., Brumfield, R.T., Faircloth, B.C., Smith, B. T., Braun, E.L., 2019. A phylogenomic supertree of birds. Diversity 11, 109. https://doi.org/10.3390/d11070109.
Kuhn, T.S., Mooers, A., Thomas, G.H., 2011. A simple polytomy resolver for dated phylogenies. Methods Ecol. Evol. 2, 427–436. https://doi.org/10.1111/j.2041-210X.2011.00103.x.
Litman, J.R., Danforth, B.N., Eardley, C.D., Praz, C.J., 2011. Why do leafcutter bees cut leaves? new insights into the early evolution of bees. Proc. Royal Soc. B 278, 3593–3600. https://doi.org/10.1098/rspb.2011.0365.
McCraney, W.T., Thacker, C.E., Alfaro, M.E., 2020. Supermatrix phylogeny resolves goby lineages and reveals unstable root of Gobiaria. Mol. Phylogenet. Evol. 151, 106862 https://doi.org/10.1016/j.ympev.2020.106862.
McFrederick, Q.S., Wcislo, W.T., Hout, M.C., Mueller, U.G., 2014. Host species and developmental stage, but not host social structure, affects bacterial community structure in socially polymorphic bees. FEMS Microbiol. Ecol. 88, 398–406. https://doi.org/10.1111/1574-6941.12302.
McGowen, M.R., Spaulding, M., Gatesy, J., 2009. Divergence date estimation and a comprehensive molecular tree of extant cetaceans. Mol. Phylogenet. Evol. 53, 891–906. https://doi.org/10.1016/j.ympev.2009.08.018.
Melin, A., Colville, J.F., 2019. A review of 250 years of South African bee taxonomy and exploration (Hymenoptera: Apoidea: Anthophila). Trans. R. Soc. South Africa 74, 86–96. https://doi.org/10.1080/0035919X.2019.1572670.

Mello, B., Schrago, C.G., 2014. Assignment of calibration information to deeper phylogenetic nodes is more effective in obtaining precise and accurate divergence time estimates. Evol. Bioinforma. 10, 79–85. https://doi.org/10.4137/EBO.S13908.
Michener, C.D., 2007. The Bees of the World. Johns Hopkins University Press, second ed., Baltimore, Maryland.
Michez, D., Patiny, S., Danforth, B.N., 2009. Phylogeny of the bee family Melittidae (Hymenoptera: Anthophila) based on combined molecular and morphological data. Syst. Entomol. 34, 574–597. https://doi.org/10.1111/j.1365-3113.2009.00479.x.
Microsoft Corporation, 2023. Microsoft Excel (Version 16.79.1).
Miller, M.A., Pfeiffer, W., Schwartz, T., 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. 2010 GCE, 1-8. https://doi.org/10.1109/GCE.2010.5676129.
Misof, B., Misof, K., 2009. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. Syst. Biol. 58, 21–34. https://doi.org/10.1093/sysbio/syp006.
Mitchell, K.J., Pratt, R.C., Watson, L.N., Gibb, G.C., Llamas, B., Kasper, M., Edson, J., Hopwood, B., Male, D., Armstrong, K.N., Meyer, M., Hofreiter, M., Austin, J., Donnellan, S.C., Lee, M.S.Y., Phillips, M.J., Cooper, A., 2014. Molecular phylogeny, biogeography, and habitat preference evolution of Marsupials. Mol. Biol. Evol. 31, 2322–2330. https://doi.org/10.1093/molbev/msu176.
Nabhan, A.R., Sarkar, I.N., 2012. The impact of taxon sampling on phylogenetic inference: A review of two decades of controversy. Brief. Bioinform. 13, 122–134. https://doi.org/10.1093/bib/bbr014.
Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274. https://doi.org/10.1093/molbev/msu300.
Oliver, P.M., Hugall, A.F., Prasteya, A., Slavenko, A., Zahirovic, S., 2023. Oligo-Miocene radiation within South-west Pacific arc terranes underpinned repeated upstream continental dispersals in pigeons (Columbiformes). Biol. J. Linn. 138, 437–452. https://doi.org/10.1093/biolinnean/blad003.
Orr, M.C., Hughes, A.C., Chesters, D., Pickering, J., Zhu, C.-D., Ascher, J.S., 2021. Global patterns and drivers of bee distribution. Curr. Biol. 31, 451–458. https://doi.org/10.1016/j.cub.2020.10.053.
Orr, M.C., Branstetter, M.G., Straka, J., Yuan, F., Leijs, R., Zhang, D., Zhou, Q., Zhu, C.-D., 2022. Phylogenomic interrogation revives an overlooked hypothesis for the early evolution of the bee family apidae (Hymenoptera: Apoidea), with a focus on the subfamily anthophorinae. Insect Syst. Divers. 6, 1–15. https://doi.org/10.1093/isd/ixac022.
Pennell, M.W., FitzJohn, R.G., Cornwell, W.K., 2016. A simple approach for maximizing the overlap of phylogenetic and comparative data. Methods Ecol. Evol. 7, 751–758. https://doi.org/10.1111/2041-210X.12517.
Peters, R.S., Meyer, B., Krogmann, L., Borner, J., Meusemann, K., Schütte, K., Niehuis, O., Misof, B., 2011. The taming of an impossible child: A standardized all-in approach to the phylogeny of Hymenoptera using public database sequences. BMC Biol. 9, 55. https://doi.org/10.1186/1741-7007-9-55.
Peters, R.S., Krogmann, L., Mayer, C., Donath, A., Gunkel, S., Meusemann, K., Kozlov, A., Podsiadlowski, L., Petersen, M., Lanfear, R., Diez, P.A., Heraty, J., Kjer, K.M., Klopfstein, S., Meier, R., Polidori, C., Schmitt, T., Liu, S., Zhou, X., Wappler, T., Rust, J., Misof, B., Niehuis, O., 2017. Evolutionary History of the Hymenoptera. Curr. Biol. 27, 1013–1018. https://doi.org/10.1016/j.cub.2017.01.027.
[dataset] Pisanty, G., Richter, R., Martin, T., Dettman, J., Cardinal, S., 2022. Molecular phylogeny, historical biogeography and revised classification of andrenine bees (Hymenoptera: Andrenidae). Mendeley Data, v5. https://doi.org/10.17632/s9mb3d437k.5.
Pisanty, G., Richter, R., Martin, T., Dettman, J., Cardinal, S., 2022b. Molecular phylogeny, historical biogeography and revised classification of andrenine bees (Hymenoptera: Andrenidae). Mol. Phylogenet. Evol. 170, 107151 https://doi.org/10.1016/j.ympev.2021.107151.
Plateaux-Quénu, C., Plateaux, L., Packer, L., 2000. Population-typical behaviours are retained when eusocial and non-eusocial forms of *Evylaeus albipes* (F.) (Hymenoptera, Halictidae) are reared simultaneously in the laboratory. Insect. Soc. 47, 263–270. https://doi.org/10.1007/PL00001713.
Potts, S.G., Imperatriz-Fonseca, V., Ngo, H.T., Aizen, M.A., Biesmeijer, J.C., Breeze, T.D., Dicks, L.V., Garibaldi, L.A., Hill, R., Settele, J., Vanbergen, A.J., 2016. Safeguarding pollinators and their values to human well-being. Nature 540, 220–229. https://doi.org/10.1038/nature20588.
Rabosky, D.L., 2015. No substitute for real data: A cautionary note on the use of phylogenies from birth-death polytomy resolvers for downstream comparative analyses. Evolution 69, 3207–3216. https://doi.org/10.1111/evo.12817.
Rabosky, D.L., Chang, J., Title, P.O., Cowman, P.F., Sallan, L., Friedman, M., Kaschner, K., Garilao, C., Near, T.J., Coll, M., Alfaro, M.E., 2018. An inverse latitudinal gradient in speciation rate for marine fishes. Nature 559, 392–395. https://doi.org/10.1038/s41586-018-0273-1.
Ramos, K.S., Martins, A.C., Melo, G.A.R., 2022. Evolution of andrenine bees reveals a long and complex history of faunal interchanges through the Americas during the Mesozoic and Cenozoic. Mol. Phylogenet. Evol. 172, 107484 https://doi.org/10.1016/j.ympev.2022.107484.
Rangel, T.F., Colwell, R.K., Graves, G.R., Fučíková, K., Rahbek, C., Diniz-Filho, J.A.F., 2015. Phylogenetic uncertainty revisited: Implications for ecological analyses. Evolution 69, 1301–1312. https://doi.org/10.1111/evo.12644.
Rees, J.A., Cranston, K., 2017. Automated assembly of a reference taxonomy for phylogenetic data synthesis. Biodivers. Data J. 5, e12581 https://doi.org/10.3897/BDJ.5.e12581.
Rehan, S.M., Chapman, T.W., Craigie, A.I., Richards, M.H., Cooper, S.J.B., Schwarz, M.P., 2010. Molecular phylogeny of the small carpenter bees (Hymenoptera: Apidae: Ceratinini) indicates early and rapid global dispersal. Mol. Phylogenet. Evol. 55, 1042–1054. https://doi.org/10.1016/j.ympev.2010.01.011.
Richards, M.H., Von Wettberg, E.J., Rutgers, A.C., 2003. A novel social polymorphism in a primitively eusocial bee. PNAS 100, 7175–7180. https://doi.org/10.1073/pnas.1030738100.
Roure, B., Baurain, D., Philippe, H., 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. 30, 197–214. https://doi.org/10.1093/molbev/mss208.
Sanderson, M.J., 2002. Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. Mol. Biol. Evol. 19, 101–109. https://doi.org/10.1093/oxfordjournals.molbev.a003974.
Sanderson, M.J., Purvis, A., Henze, C., 1998. Phylogenetic supertrees: Assembling the trees of life. Trends Ecol. Evol. 13, 105–109. https://doi.org/10.1016/S0169-5347(97)01242-1.
Sanderson, M.J., Mcmahon, M.M., Stamatakis, A., Zwickl, D.J., Steel, M., 2015. Impacts of terraces on phylogenetic inference. Syst. Biol. 64, 709–726. https://doi.org/10.1093/sysbio/syv024.
Sanderson, M.J., McMahon, M.M., Steel, M., 2010. Phylogenomics with incomplete taxon coverage: the limits to inference. BMC Evol. Biol. 10, 155. https://doi.org/10.1186/1471-2148-10-155.
Sann, M., Niehuis, O., Peters, R.S., Mayer, C., Kozlov, A., Podsiadlowski, L., Bank, S., Meusemann, K., Misof, B., Bleidorn, C., Ohl, M., 2018. Phylogenomic analysis of Apoidea sheds new light on the sister group of bees. BMC Evol. Biol. 18, 71. https://doi.org/10.1186/s12862-018-1155-8.
Sann, M., Meusemann, K., Niehuis, O., Escalona, H.E., Mokrousov, M., Ohl, M., Pauli, T., Schmid-Egger, C., 2021. Reanalysis of the apoid wasp phylogeny with additional taxa and sequence data confirms the placement of Ammoplanidae as sister to bees. Syst. Entomol. 46, 558–569. https://doi.org/10.1111/syen.12475.
Schwarz, M.P., Fuller, S., Tierney, S.M., Cooper, S.J.B., 2006. Molecular phylogenetics of the exoneurine allodapine bees reveal an ancient and puzzling dispersal from Africa to Australia. Syst. Biol. 55, 31–45. https://doi.org/10.1080/10635150500431148.
Shen, X.X., Li, Y., Hittinger, C.T., Chen, X., Rokas, A., 2020. An investigation of irreproducibility in maximum likelihood phylogenetic inference. Nat. Commun. 11, 6096. https://doi.org/10.1038/s41467-020-20005-6.
[dataset] Sless, T., 2021. Phylogenetic relationships and the evolution of host preferences in the largest clade of brood parasitic bees (Apidae: Nomadinae). figshare, v1. https://doi.org/10.6084/m9.figshare.c.5556573.v1.
Sless, T.J.L., Branstetter, M.G., Gillung, J.P., Krichilsky, E.A., Tobin, K.B., Straka, J., Rozen, J.G., Freitas, F.V., Martins, A.C., Bossert, S., Searle, J.B., Danforth, B.N., 2022. Phylogenetic relationships and the evolution of host preferences in the largest clade of brood parasitic bees (Apidae: Nomadinae). Mol. Phylogenet. Evol. 166, 107326. https://doi.org/10.1016/j.ympev.2021.107326.
Smirnov, V., Warnow, T., 2021. Phylogeny estimation given sequence length heterogeneity. Syst. Biol. 70, 268–282. https://doi.org/10.1093/sysbio/syaa058.
Smith, M.R., 2022. Using information theory to detect rogue taxa and improve consensus trees. Syst. Biol. 71, 1088–1094. https://doi.org/10.1093/sysbio/syab099.
Smith, S.A., Brown, J.W., 2018. Constructing a broadly inclusive seed plant phylogeny. Am. J. Bot. 105, 302–314. https://doi.org/10.1002/ajb2.1019.
Smith, S.A., O'Meara, B.C., 2012. treePL: Divergence time estimation using penalized likelihood for large phylogenies. Bioinformatics 28, 2689–2690. https://doi.org/10.1093/bioinformatics/bts492.
Soltis, P.S., Soltis, D.E., Wolf, P.G., Nickrent, D.L., Chaw, S.M., Chapman, R.L., 1999. The phylogeny of land plants inferred from 18S rDNA sequences: Pushing the limits of rDNA signal? Mol. Biol. Evol. 16, 1774–1784. https://doi.org/10.1093/oxfordjournals.molbev.a026089.
Stamatakis, A., 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033.
Stork, N.E., 2018. How many species of insects and other terrestrial arthropods are there on earth? Annu. Rev. Entomol. 63, 31–45. https://doi.org/10.1146/annurev-ento-020117-043348.
Thomas, G.H., Hartmann, K., Jetz, W., Joy, J.B., Mimoto, A., Mooers, A.O., 2013. PASTIS: An R package to facilitate phylogenetic assembly with soft taxonomic inferences. Methods Ecol. Evol. 4, 1011–1017. https://doi.org/10.1111/2041-210X.12117.
Upham, N.S., Esselstyn, J.A., Jetz, W., 2019. Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. PLoS Biol. 17, e3000494 https://doi.org/10.1371/journal.pbio.3000494.
Wagner, D.L., 2020. Insect declines in the anthropocene. Annu. Rev. Entomol. 65, 457–480. https://doi.org/10.1146/annurev-ento-011019-025151.
Wiens, J.J., Morrill, M.C., 2011. Missing data in phylogenetic analysis: Reconciling results from simulations and empirical data. Syst. Biol. 60, 719–731. https://doi.org/10.1093/sysbio/syr025.
Winfree, R., Aguilar, R., Vázquez, D.P., LeBuhn, G., Aizen, M.A., 2009. A meta-analysis of bees' responses to anthropogenic disturbance. Ecology 90, 2068–2076. https://doi.org/10.1890/08-1245.1.
Yang, Z., Rannala, B., 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol. Biol. Evol. 23, 212–226. https://doi.org/10.1093/molbev/msj024.
Zheng, Y., Wiens, J.J., 2015. Do missing data influence the accuracy of divergence-time estimation with BEAST? Mol. Phylogenet. Evol. 85, 41–49. https://doi.org/10.1016/j.ympev.2015.02.002.