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

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

Affiliation

[a] School of BioSciences, The University of Melbourne, Parkville, Victoria, Australia

[b] Department of Sciences, Museums Victoria, GPO Box 666, Melbourne, Victoria, Australia

*A.F.H and P.H.-P. contributed equally to this work.

Corresponding authors:

Andrew F. Hugall[a b]

ahugall@museum.vic.gov.au

Patricia Henríquez-Piskulich[a]

henriquezpiskulichp@gmail.com

Highlights

- Bee supermatrix phylogeny constructed with public and published sequence data.

- Includes 23% of currently recognised species and covers 86% of genera.

- Provides a summary of remaining gaps in bee phylogenetics.

- Available online at beetreeoflife.org, with subsetting tool to facilitate comparative analyses.

Abstract

The increasing availability of large phylogenies has provided new opportunities to study the evolution of species traits, their origins and diversification, and biogeography; yet, with the exception of butterflies, taxonomically well-curated phylogenies are currently lacking 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 relationships between and/or within families or genera for this group have been published. We used publicly available sequence data, a family-level phylogenomic backbone, and ultra-conserved element (UCE) data, reconciled to a taxonomic database, to produce a dated phylogeny for bees. The phylogeny comprises 4651 bee species, representing 23% of species and 86% of genera. At family, subfamily, and tribe levels, the data were robust, but between and within some genera relationships remain uncertain. In addition, most of the species with available sequence data are geographically distributed in North America and Europe, highlighting gaps that should be considered in future research to improve our understanding of bee evolution and phylogeography. We provide a summary of the current state of molecular data available and its gaps, and 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.

Keywords

Bees; molecular phylogenetics; systematics; macroevolution

1. Introduction

Phylogenies are a key tool, not only for evolutionary biology, but also ecology subdisciplines such as conservation biology, community ecology and macroecology (Beck et al., 2012; Harvey and Pagel, 1991; Mouquet et al., 2012). In the last two decades, phylogenetic supermatrices of several groups have become available, opening new opportunities to study the evolution of species traits, their origins and 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 (Mora et al., 2011), well-curated large scale phylogenies are currently lacking for major insect groups, with the exception of butterflies (Adler and Foottit, 2018; Chazot et al., 2021; Kawahara et al., 2023).

Bees (Hymenoptera: Anthophila) are a large and diverse group of insect pollinators with over 20000 described species and a worldwide distribution (Orr et al., 2020). This group has received increased attention in recent decades given their importance as pollinators and the future consequences 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 and do not build a nest or forage for food and instead 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 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 attempt to produce a supermatrix phylogeny of bees was by Hedtke et al. (2013) a decade ago. Since then, the amount of data that has become available has increased dramatically, demanding

a substantially updated phylogeny for this group. Numerous molecular phylogenetic studies of relationships between and/or within families or genera for this group have been published over the years (Supplementary Table S1). In addition, comprehensive molecular phylogenies have been published for the seven bee families (Melittidae, Andrenidae, Halictidae, Colletidae, Stenotritidae, Megachilidae and Apidae (Michener, 2007)). Although some uncertainties remain unresolved, the estimates of most of their evolutionary relationships between families and subfamilies are robust (Almeida et al., 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). As a result of this volume of phylogenetic studies, a substantial amount of phylogenetic data has become available for bees which can be used to build a supermatrix phylogeny.

Here, we use publicly available sequence data, a family-level transcriptome phylogenomic backbone, and ultra-conserved elements (UCE), reconciled to a taxonomic database to ensure taxonomic consistency, to construct a supermatrix phylogeny of bees. 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. Once all components were curated, we used them to produce a concatenated supermatrix to construct maximum likelihood trees. This approach produced the largest bee phylogeny to date that includes 4651 species in 427 genera. While this supermatrix phylogeny doesn't aim to solve current uncertainties in phylogenetic relationships, it is a curated composite of extensive previous phylogenetic work 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 1000 bootstrap sample trees converted to dated chronograms to represent the phylogenetic information and uncertainty in our 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, a taxonomic database was developed for Anthophila based on data provided by Orr et al. (2020), 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. This 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 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; Jetz et al., 2012; Upham et al., 2019) to build a large bee 'tree-of-life' over other methods (e.g., supertree: Bininda-Emonds, 2004; Sanderson et al., 1998). Supermatrix methods use all available information to estimate relationships which provides higher resolution, simultaneously providing branch length and uncertainty (bootstrap resampling). We built our phylogenetic supermatrix database for bees based on four key components: i. gene fragment accessions downloaded from global sequence databases; ii. a family-level backbone using a subset of transcriptome phylogenomic data; iii. published UCE datasets combined and condensed; and iv. nomenclature reconciled to the taxonomic database.

Generally, efficient supermatrices have a hierarchical mix of 'fast' and 'slow' genes, which typically involves a mitochondrial gene for the tips and a set of nuclear loci that can be limited to representatives of the major lineages (Beaulieu and O'Meara, 2018; Driskell et al., 2004). Thus, instead of sheer data volume, we focussed on fewer widely sampled loci used in published studies with the greatest proportion of taxa and phylogenetic spread to maximize overlap of data among lineages. This approach can make it easier to check individual genes, and limit the proportion of taxa without data in common, which can distort phylogenetic inference (Freyman, 2015; Sanderson et al., 2010; Wiens and Morrill, 2011). Thus, the publicly available gene fragment accessions provided the species (tips), the transcriptome phylogenomic data provided a robust backbone consistent with phylogenomic information on tree shape (topology, relative branch length and outgroup rooting), and the UCE leveraged new and powerful phylogenomic data that is increasingly being used, to further enforce best-estimate tree shape. Collation and curation of each component was done with custom Bash

124 scripts and Microsoft Excel (see Supplementary file 2 and Supplementary file 3) (Microsoft
125 Corporation, 2021); details for each component are given in the following sections. Given the multiple
126 data sources, bespoke taxonomic database, and the need for hands-on refinement, we expanded our
127 supermatrix assembly system used in previous works (e.g., Hugall and Stuart-Fox, 2012; Oliver et al.,
128 2023).
129
130 **2.2.1. Gene fragment accessions**
131 Data were downloaded from the NCBI nucleotide collection (https://www.ncbi.nlm.nih.gov/genbank).
132 We used gene sequences obtained from key phylogenetic references in BLAST+organism searches
133 with an E-value threshold of 5e-6 and retained aligned regions, excluding model and environmental
134 data. For protein coding genes and ribosomal RNA genes we used tblastn and blastn respectively. In
135 total, we obtained data for seven nuclear protein-coding genes (ArgK, CAD, NaK, Pol II, Wnt-1, LW Rh,
136 and EF-1α), two ribosomal genes (16S and 28S), and two mitochondrial coding genes (Cytb and COI)
137 (Table 1).
138
139 **Table 1**. Sampling summary across genetic data.

Gene*	Sites	Genera	Species	p-sp	Description
Nuclear					
Phylogenomic 'stub'	12440	36	40	0.01	Peters et al. 2017 exon data subset (1/2 codon positions)
UCE 'stub'	13250	179	678	0.15	Composite UCE data subset (Bossert, Sless, Freitas, and Pisanty)
ArgK	546	63	613	0.13	Arginine kinase gene, exon 2
CAD	1014	138	682	0.15	Conserved ATPase domain protein gene exon
EF-1α	1113	332	2047	0.44	Elongation factor 1-alpha gene exon
LW Rh	684	316	1848	0.40	Long wavelength sensitive opsin 1 gene exon
NaK	1461	231	790	0.17	Sodium potassium adenosine triphosphatase gene exon
Pol II	840	198	853	0.18	RNA polymerase II gene exon
Wnt-1	456	255	843	0.18	Wingless gene exon
28S rDNA	1464	339	1257	0.27	Large subunit nuclear ribosomal RNA gene
Mitochondrial					
16S rDNA	536	76	513	0.11	Mitochondrial large subunit ribosomal RNA gene
COI	1476	327	3892	0.84	Mitochondrial cytochrome oxidase subunit 1 gene
Cytb	1047	65	487	0.11	Mitochondrial cytochrome b gene
Total	36327	2555	14543		

*For our purposes, 'gene' refers to a discrete genetic data unit.
p-sp: proportion of species with gene.

140
141 We focused on protein coding regions (e.g., exons), as they are easier to align, and make it simple to
142 add new data. For the purposes of this work, we did not include the widely sequenced ribosomal 18S
143 gene due to its low and poor phylogenetic signal (Soltis et al., 1999). Because of the limited coverage
144 of whole mitogenome data for bees (Husemann et al., 2021), we extracted only near full-length
145 mitochondrial genes (16S, Cytb, COI) with the reference BLAST. Rather than selecting one 'best'

146 longest species accession per locus, we retained all accessions to first assess the gene tree monophyly
147 of nominate species as well as genera. Due to the great number of accessions, EF-1α, LW Rh and COI
148 genes were split into two sets: Apidae and the rest.
149
150 Gene accessions were aligned with MAFFT v. 7.245 (Katoh and Standley, 2013) using the auto setting.
151 We then assembled alignments into a custom Microsoft Excel database, and reconciled nomenclature
152 with the taxonomic list (Supplementary file 1). We refined protein coding gene alignments by hand in
153 BioEdit (Hall, 1999) using the original references as a guide, to have all coding regions strictly in-frame.
154 Despite being time consuming, aligning protein coding genes this way makes adding taxa simpler. For
155 the ribosomal RNA genes, we used the MAFFT alignment as is.
156
157 We inferred gene trees through IQ-TREE (Nguyen et al., 2015), and in some cases RAxML (Stamatakis,
158 2014). IQ-TREE used ultrafast bootstrap consensus (Hoang et al., 2018) with models of sequence
159 evolution identified by ModelFinder implemented in IQ-TREE (Kalyaanamoorthy et al., 2017). RAxML
160 used fast bootstrap and the GTR+G model, with the standard -f a setting implemented on CIPRES
161 (Miller et al., 2010). We assessed gene trees for non-monophyletic genera and species, and aberrant
162 sequences such as paralogs (e.g., EF-1α and LW Rh variants) and gross miss-alignments (long terminal
163 branches). Then, we updated the dataset by removing abnormal accessions and in some cases revising
164 the nomenclature. Datasets were realigned if necessary.
165
166 The mitochondrial COI was processed slightly differently because of the extensive number of
167 accessions that are publicly available, its nominal intra-specific diversity, and because it is often
168 sequenced in separate fragments. We reduced this gene alignment to a single consensus sequence
169 per species, based on the most common base per site (with ties scored as ambiguous). This approach
170 tends towards the most commonly sequenced sub-lineage and is a simple way to combine data,
171 discount rarer aberrant sequences and rationalize choice of intra-specific lineage complexity. The
172 consensus alignment was then subjected to the same procedure of gene tree and genera monophyly
173 assessment as described for the other genes included in this work. Additionally, we included COI data
174 from The Barcode of Life Data System (BOLD, http://www.boldsystems.org) for a selection of species
175 that had multiple samples that clustered closely (>5%), or that represented taxa we already had with
176 other genes if there was only one sequence available, and that fell within the lowest rank (genus or
177 subfamily depending on taxon sampling) in our working COI tree. Information on data sources for all
178 taxa and genes is summarized in Supplementary file 3.
179

180 **2.2.2. Family-level phylogenomic backbone**

181 Hymenoptera transcriptome phylogenomic work of Peters et al. (2017) provided the current best
182 estimate of the higher family level phylogenetic tree for Anthophila. This work included data for 41
183 bee taxa from six of the seven families (excepting Stenotritidae) for 3256 genes. This vast amount of
184 data was unnecessary for the purposes of the bee supermatrix and thus, we extracted a subset
185 sufficient to reconstruct tree shape. To this aim, we trimmed down the first and second position exon
186 datamatrix from [dataset] Peters et al. (2017) to the Anthophila and key closest 'Crabronidae'
187 outgroups (60 taxa), and condensed codons (3.01 million bases, 1.5 million codons) down to a practical
188 amount (12.4 kb). This was done by deleting codons with <85% of taxa and then randomly sub-
189 sampling one third of codons (jack-knife), thereby reducing the proportion of missing data from the
190 original 57.3% complete to 92.8% complete. We then used this phylogenomic 'stub' to produce a tree,
191 as previously described for the gene fragment datasets, and compared it with the original
192 Hymenoptera phylogeny by Peters et al. (2017).

193

194 **2.2.3. UCE data**

195 Ultra-conserved elements are a method of extracting phylogenomic scale data that is increasingly
196 being used for bees. Because this method produces more data than is necessary for the bee
197 supermatrix phylogeny, we selected and combined a subset of data from five previously published
198 data-matrices ([dataset] Bossert, 2021a, 2021b; [dataset] Freitas et al., 2020; [dataset] Pisanty et al.,
199 2022; [dataset] Sless, 2021). First, we produced consensus sequences for each of the five data-
200 matrices with the same approach described for the COI gene. Then the 1388 UCE consensus loci
201 obtained from [dataset] Bossert (2021a) were mapped onto the consensus sequences of the other
202 data-matrices with BLAST v2.13.0, to identify the corresponding orthologous sections. We then
203 selected a subset of suitable UCE loci using several criteria: maximum e-value of 1e-50, >150 sites, and
204 >50% sites matched to the consensus reference. We aligned these loci with MAFFT as previously
205 described for gene fragments, concatenated them into a UCE 'supermatrix', and reconciled species
206 names with the taxonomic database. This supermatrix (57.9 kb in 94 loci for 777 samples) was then
207 compacted down to an amount adequate to reconstruct key tree shape, without excess computational
208 burden, by deleting sites that were <74% complete and then further reduced by one third jack-knifing,
209 resulting in a dataset of 13.3 kb; 85% complete. We then produced a tree to compare to the original
210 published UCE phylogenies, and to assess for non-monophyletic genera and aberrant sequences as
211 previously described. The resulting tree had 71% of nodes with bootstrap support >=90%. Details of
212 the primary transcriptome and UCE data sources are provided in Supplementary file 3.

213 There are three approaches when working with UCE sequence data: i. Reconstruct from primary reads;
214 ii. reconstruct from locus datasets; iii. reconstruct from the final processed and refined supermatrix.
215 We took the third approach, as a simple compromise that has the benefit of leveraging from previous
216 bioinformatic work. We regard it preferable to 'supertree' approaches (e.g., Kimball et al., 2019) that
217 merely use the topology of phylogenomic datasets, with branch length/ages applied post-hoc.
218
219 **2.3. Supermatrix assembly and analysis**
220 Once all supermatrix data components were reconciled to the taxonomic database, aligned and
221 assessed for aberrant elements, they were used to produce a concatenated supermatrix. We did this
222 by taking the best (longest accepted) single exemplar sequence per 'gene' per species, for all species
223 with a total of least 400 sites of any data. The supermatrix was then compacted by removing regions
224 with little or no data (<10% taxa with any data per 'gene') and ambiguous alignment regions with
225 Aliscore v2.2 (Misof and Misof, 2009), while maintaining protein coding reading frames.
226
227 Sparse, patchy supermatrices often include taxa groups that share no data in common, which can bias
228 the phylogenetic inference (Sanderson et al., 2010; Smirnov and Warnow, 2021; Wiens and Morrill,
229 2011). To ameliorate this issue, we implemented minimal taxonomic rank substitution to improve data
230 density on the phylogenomic 'stub', and species with no overlap of data within genera (Pennell et al.,
231 2016). For genera that contained species that did not share any data with any other member (i.e., the
232 only species in that rank with that genetic data), we reassigned the data to the member of the genus
233 with the most data and removed the source species. In total 40 substitutions were made (0.4% of the
234 matrix).
235
236 The outgroup included five taxa representing the Crabronidae subfamilies Astatinae, Bembicinae,
237 Crabroninae and Philanthinae, based on the phylogenomic 'stub' plus a small amount of gene
238 fragment data. As gene data was patchy across outgroups, taxonomic rank-substitution was also used
239 to create composite family representatives for some taxa. The gene composites included the genera
240 *Astata* (Astatinae), *Bembix* (Bembicinae), *Cerceris* (Philanthinae), *Philanthus* (Philanthinae) and
241 *Tachysphex* (Crabroninae).
242
243 To reduce computational burden, the supermatrix was split into three subsets comprising the families
244 Megachilidae + Andrenidae + Melittidae, Halictidae + Colletidae + Stenotritidae, and Apidae, each
245 sharing a core set of 18 family representatives with maximal data, and the five outgroups. In addition,
246 we extracted one species with the most data for each nominal genus to create a genus-level

247 supermatrix. For each of these supermatrix subsets, phylogenetic inference was done by IQ-TREE with
248 ultrafast bootstrap, using a RAxML starting tree, all nearest-neighbour interchange and refined tree
249 search settings (-allnni, -beps 8, -bnni, -bcor 0.98, -wbtl, -pers 0.4, -nstop 200, --sprrad 8; via CIPRES).
250 The intention was to focus effort on tree support space, in what can be a difficult problem (Shen et
251 al., 2020). The optimal partitioned sequence evolution model was determined with ModelFinder (in
252 IQ-Tree) using the genus-level supermatrix, and this 8-partition model scheme was then used for all
253 remaining analyses (see Supplementary file 3). This ensures a common model irrespective of the
254 amount of data per gene per family subset (i.e., as if all three were analysed simultaneously). The trees
255 from the three family subsets were then joined together again using the common family
256 representatives, for both the consensus tree and the 1000 bootstrap trees.
257
258 **2.4. Tree dating calibration**
259 Patchy and missing data can exacerbate unequal/uneven estimates of branch length among lineages
260 (how clock-like the tree appears), over and above inherent rate variation (Roure et al., 2013; Zheng
261 and Wiens, 2015). To reduce this calibration problem, it is necessary to rely on common partition
262 models, and relaxed-clock models. The species-level supermatrices were far too large to properly run
263 Bayesian relaxed-clock methods (Fisher et al., 2022), and the divide and graft approach is complicated
264 and constraining (e.g., Jetz et al., 2012; Upham et al., 2019). Thus, to overcome these issues we sought
265 a simpler approach, explained in the following paragraphs, that balanced computational effort with
266 the limited precision of the data.

267 Trees were converted into chronograms using Penalised Likelihood Rate Smoothing (PLRS)
268 (Sanderson, 2002), using treePL (Smith and O'Meara, 2012). To enforce chronograms to remain fully
269 bifurcating and improve the efficiency of the PLRS iteration, a small length increment was added to
270 any branch of length >0.0005 (amounting to mean 0.43% of the original total tree length, affecting
271 6.8% of branches). TreePL was run with 'thorough prime' optimized settings and smoothing factor=10.
272 Because the all-species tree was too large to use cross-validation, a suitable smoothing factor was
273 empirically determined from the genus-level tree.

274 There are few certain dating calibrations for major lineages within bees and considerable variation
275 among published works (e.g., Almeida et al., 2012; Cardinal and Danforth, 2013; Freitas et al., 2022;
276 Peters et al., 2017; Schwarz et al., 2006). Many studies rely on secondary dating from a limited number
277 of previous works (e.g., Martins et al., 2014; Rehan and Schwarz, 2015; Trunz et al., 2016). Therefore,
278 rather than enforce a range of putative calibrations, we simply calibrated the bee root to a broad
279 normal distribution of 100 million years ago (mya) SD 4. This was achieved by drawing a root age from
280 a normal distribution for each of the 1000 bootstrap replicates. This was used for both the genus-level

and the all-species trees. The results were then compared to published ages/calibrations 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 is 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. This phylogeny is not a test of bee systematics but a reflection of current information and ideas about bee relationships. 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 4651 species and 427 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 36327 sites with a total of 18.6 million base characters from 14543 data elements, median 11% complete (3981 sites 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 April 2023, providing the largest supermatrix dataset and tree of bees to date (Chesters, 2020; 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. Between families, sampling coverage of named genera ranged from 73- 100% but at the species level, coverage was lower ranging between 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 our 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

315 other regions (see Supplementary file 4 for a description of the methods used to obtain bee
316 distribution data). Despite bees being a diverse insect group with global distribution (Danforth et al.,
317 2019; Michener, 2007), geographical biases remain. This has previously been discussed in terms of
318 ecological gaps (Archer et al., 2014), and taxonomic gaps (Orr et al., 2020), but efforts should also be
319 made to obtain molecular data from other regions to better understand the relationships and the
320 evolution of bees, especially considering that some of these less studied regions might host high bee
321 species richness (Batley and Hogendoorn, 2009; Eardley et al., 2009; Freitas et al., 2009; Groom and
322 Schwarz, 2011; Melin and Colville, 2019).
323
324 **Table 2**. Summary of taxonomic sampling by family.

Family	Total subfamilies*	Total genera	Included genera (%)	Total species	Included species (%)
Andrenidae	3	50	43 (86)	2957	644 (22)
Apidae	5	211	184 (87)	5829	1779 (31)
Colletidae	9	76	64 (84)	2616	523 (20)
Halictidae	4	82	62 (76)	4400	779 (18)
Megachilidae	3	83	61 (73)	4099	838 (20)
Melittidae	3	15	11 (73)	201	82 (41)
Stenotritidae	1	2	2 (100)	21	6 (29)
Total	28	519	427 (82)	20123	4651 (23)

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

325
326 **3.2. Gene sampling**
327 Overall, the data were highly skewed with many species (38%) having one gene and only a small
328 number (16%) having more than five (Fig. 3A-B). However, data were well distributed across linages
329 and a small number of species with many genes were phylogenetically spread across the tree (Fig. 3A-
330 C). In addition, the use of shortened subsets of the very large phylogenomic data (transcriptome and
331 UCE) provided sufficient information to ensure the tree was consistent with published higher-level
332 phylogenies in topology and relative branch length. Our phylogenomic 'stub' recreated published
333 phylogenomic trees with high bootstrap support (>95%), and the UCE ensured that most of the key
334 nodes matched previously published results (see Similarity to previous studies section). Thus, the
335 addition of thousands of taxa with little data to the core backbone had limited effect on the underlying
336 backbone of the tree. This can be seen in the similarity of the genus representative tree and the
337 equivalent subtree pruned from the full all-species tree where 89% of nodes were recovered;
338 remaining nodes had low support in either or both trees (Fig. S1). Branch lengths were also very similar
339 as are the inferred clade ages (see section below).
340

341 Based on cumulative IQTree ultrafast bootstrap support values (ufbs) (Fig. 3D), median ufbs support
342 was 89, while 23% of nodes were below 70. This varied across the family subtrees. The Apidae and
343 Melittidae had the best support (median ufbs 94 for both), and Colletidae the least (median 77). The
344 genus-level representative tree returned higher support: median ufbs 95 (Fig. 3D, Fig.S1).
345
346 **3.3. Data structure limitations**
347 A key issue in supermatrix data structure is how much data is shared between true sister taxa, a simple
348 measure being the genes taxa have in common (hereafter referred to as 'GIC' – genes in common).
349 GIC is very similar to taxon triplets measures (Sanderson et al., 2010), and provide an indication of
350 data structure limitations by identifying groups that are most affected by scarce data overlap, and
351 would therefore profit from targeted data gap filling (see Freyman, 2015). In a cladistic analysis, taxa
352 with no data in common cannot be recovered as sister taxa (Sanderson et al., 2010). In principle, a
353 supermatrix would have at least one GIC to all species (typically mtDNA) with the remaining genes
354 (typically nuclear) limited to key representatives of the major lineages. However, in practice
355 supermatrices tend to be patchy with much missing data, potentially leaving many true sister taxa
356 with little or no data in common. Nonetheless, our most common gene COI covers 3892 species (84%,
357 Table 1). The nuclear genes EF-1a and LW Rh are also widely sampled (44 and 40%, respectively).
358
359 At family and subfamily levels, data was highly robust and structured with a minimum of seven GIC
360 (median of eight) among all subfamilies (Supplementary Table S5, Fig. 3C). Within genera the overlap
361 of data varied (Supplementary Table S5, Fig. S2A-B), with numerous cases of zero shared data between
362 individuals within a genus (see Supplementary Table S6 for details). The generally increasing GIC
363 deeper into the tree and variation in GIC towards the tips is apparent from GIC mapped onto the tree
364 (Figure S2C; note that in the tree GIC should be >0). There are two approaches when dealing with
365 limited GIC, either remove species or fill in the existing gaps. Excluding rogue taxa may also help but
366 this can be a complicated process (Smith, 2022). Until these gaps are filled in, species relationships
367 within poorly structured genera should be treated with caution, and in the cases where analyses are
368 sensitive to tree topology, these genera could be pruned back to single genus representatives.
369
370 **3.4. Similarity to previous studies**
371 As intended, our tree largely reproduces the recent best phylogenomic scale results for families and
372 subfamilies (e.g., Bossert et al., 2021, 2019; Branstetter et al., 2017; Peters et al., 2017; Sann et al.,
373 2018; Sless et al., 2022). These differ somewhat to older multi-gene studies, the data of which we also
374 included (Almeida et al., 2012; Cardinal et al., 2018; Cardinal and Danforth, 2013; Ramos et al., 2022;

375 Rehan et al., 2010). There are some differences between our full species tree and our genus
376 representative tree, and these differences mirror the uncertainties reported in some of the primary
377 studies. Most of these differences involve short internodes between lineages. Even with massive
378 amounts of multi-locus UCE data such situations can be unresolvable and/or uninformative (Degnan
379 and Rosenberg, 2009; Hahn and Nakhleh, 2016). Some key remaining uncertainties include the
380 subgeneric relationships within *Andrena* (Andrenidae: Andreninae) given its high degree of paraphyly
381 and polyphyly, where classification changes will likely occur with further efforts to resolve such a
382 diverse and complex genus (Pisanty et al., 2022). Similarly for the enormous genus *Lasioglossum*
383 (Halictidae: Halictinae). The position of the subfamily Anthophorinae (Apidae) is also uncertain, having
384 been recovered as sister to Nomadinae (Bossert et al., 2019), as well as sister to all remaining Apidae
385 families (Orr et al., 2022). Our all-species tree agrees with the latter but with weak support (ufbs 73%)
386 while our genus tree has yet another result (Fig.S1). A similar case applies with the mono- or paraphyly
387 of the Centridini. We summarise key differences between this phylogeny and previously published
388 work in Supplementary file 5, which also mentions some genera that are potentially not monophyletic.
389

390 **3.5. Dating: keeping it simple**

391 For large phylogenies, Bayesian dating methods are computationally intensive. Penalised likelihood
392 rate smoothing (PLRS) as implemented in TreePL, is an efficient alternative without the complexities
393 of the divide and graft approach (e.g., Jetz et al., 2012; Smirnov and Warnow, 2021; Upham et al.,
394 2019). We took the simplest dating approach by allowing the data and relaxed-clock model to define
395 the relative age pattern, applying a single overall root calibration to scale the absolute age, the single
396 most important calibration in relaxed-clock dating (Kawahara et al., 2023; Mello and Schrago, 2014;
397 Sanderson, 2002). This result was then used to compare how well the broad tree shape matches
398 previous dating efforts and proposed fossil-based internal calibrations (Beaulieu and O'Meara, 2018;
399 Hugall et al., 2007; Near et al., 2005).
400

401 We compared a range of groups across families, for which there were multiple previous estimates
402 from 19 different published studies (Fig. S3A; Supplementary Table S7). We also looked at the 34 fossil
403 calibrations in Cardinal et al. (2018), including 15 previously described in detail in Cardinal and
404 Danforth (2013) (Fig. S3B). Our dating falls within the range for most taxa (Fig. S3A), representing a
405 broad consensus framework, and the all-species and genus only tree dates are highly correlated
406 (0.994). Most of our dates fall either within the Cardinal et al. (2018) prior 95% confidence intervals
407 (CI) or above their minimum age constraint (Fig. S3B). These comparisons highlight the limited
408 knowledge of the age of bees as a whole or the age of many of the lineages. In addition, many studies

409 use previous results as secondary calibrations, meaning there is even less known than it appears.
410 General consilience of tree and fossil constraints suggest that the data+model is reasonable for
411 estimating a time tree. The unconstrained tree is sufficiently clock-like that apparent rate variation
412 can be accommodated by the PLRS correlated model (e.g., Fig. S1).
413
414 The fossil constraints are either much younger than the lineage appears to be, or they are close. As
415 they are essentially minimum age constraints (and would only be applied as such in PLRS), they would
416 have limited influence (Kawahara et al., 2023). There are a few exceptions worth noting where we
417 found significant disparities to the Cardinal et al. (2018) fossil calibrations (Table 3, Fig. S3B).
418
419 **Table 3**. Differences with Cardinal et al. (2018) fossil calibrations.

	Notes
Andrenidae 14, 19, 20	Dates for Andreninae and Pangurninae are older, but our taxon sampling is much greater and our Andrenidae dates fall within the range of previous studies (Fig. S3A). Cardinal et al. (2018) dates are at or outside their prior maximum CI.
Apidae 3, 24	Given that the genus *Apis* is not sister to the Euglossini tribe (Bossert et al., 2019), the stem for Apidae might be deeper. In Cardinal et al. (2018) the dating for Apidae is at the maximum of their prior CI. In the case of Melectini, our dating is older but in Cardinal et al. (2018) posterior is at their prior CI maximum.
Apidae 4, 5	These involve the stingless bees (Meliponini). Our dates are younger, and the larger taxon sampling reveals the complexity of genera involved. The whole group appears to have a shallow crown and a long stem. This could be the result of using fossils that are stem taxa and not extant crown taxa, an uncertainty posited for many of the Cardinal et al. (2018) calibrations.
Melittidae 33	Our date for this family is older but considering it is sister to all other families there is scope for a deeper age.
Root (crown age) of the extant Anthophila 17	Cardinal et al. (2018) based this on the fossil *Melittosphex burmensis* but a recent paper suggests it might not be a bee (Rosa and Melo, 2021). Previous estimates, including large hymenopteran phylogenomic studies, returned a wide range of ages. Thus, we kept to a broad consensus normal distribution age (mean 100 mya, SD 4) for the extant Anthophila, minimizing the disparity.

420
421 **3.6. Direction of future efforts**
422 Considering the amount of UCE data that has become available in recent years, we anticipate large
423 projects that will consolidate all primary UCE. With these data it would be possible to build a more
424 comprehensive bee backbone 'stub'. In the case of data structure commonality, a targeted approach
425 would be the best way to overcome the current gaps such as the recent targeted exon capture method
426 used to build a global phylogeny for butterflies (Kawahara et al., 2023). Gaps in bee molecular data
427 that emerged from this work are outlined in our summaries of data patchiness. At present most genera
428 are represented but within genera, species sampling is uneven, ranging between zero to 100% but
429 often below 20%. A few key 'legacy genes' would be needed, at least initially, such as COI and EF-1α.
430 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 for the use of approaches that deal with missing data (Garamszegi and Møller, 2011), and the importance of always assessing the assumptions and biases of phylogenetic comparative analyses, as to avoid poor model fits and misinterpretation of results (Cooper et al., 2016; Rangel et al., 2015). Future efforts should be placed in 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 extensive bee phylogeny to date, providing a summary of substantial existing sequence data available up to April 2023, as well as its gaps. The supermatrix phylogeny was built with public data curated with some care, 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 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.

Acknowledgments

Patricia Henríquez-Piskulich was supported by funding awarded from the Agencia Nacional de Investigación y Desarrollo de Chile (Scholarship ID 72210037).

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).

References

[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] 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

[dataset] Peters, R., Krogmann, L., Mayer, C., Donath, A., Gunkel, S., Meusemann, K., Kozlov, A., Podsiadlowski, L., Petersen, M., Lanfear, R., Dietz, P., Heraty, J., Kjer, K., 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. Mendeley Data, v3. https://doi.org/10.17632/trbj94zm2n.3

[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

[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

Adler, P.H., Foottit, R.G., 2018. Insect biodiversity: science and society. John Wiley & Sons.

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., 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

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

Beck, J., Ballesteros-Mejia, L., Buchmann, C.M., Dengler, J., Fritz, S.A., Gruber, B., Hof, C., Jansen, F., Knapp, S., Kreft, H., Schneider, A.K., Winter, M., Dormann, C.F., 2012. What's on the horizon for macroecology? Ecography (Cop.). 35, 673–683. https://doi.org/10.1111/j.1600-0587.2012.07364.x

Bininda-Emonds, O.R.P., 2004. The evolution of supertrees. Trends Ecol. Evol. 19, 315–322. https://doi.org/10.1016/j.tree.2004.03.015

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 (N. Y). 72, 590–600. https://doi.org/10.1111/evo.13446

Cardinal, S., Danforth, B.N., 2013. Bees diversified in the age of eudicots. Proc. R. Soc. B Biol. Sci. 280. 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 Practical Solutions 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, 1–10. https://doi.org/10.1038/s41467-021-25906-8

559 Chesters, D., 2020. The phylogeny of insects in the data-driven era. Syst. Entomol. 45, 540–551.
560 https://doi.org/10.1111/syen.12414

561 Cooper, N., Thomas, G.H., FitzJohn, R.G., 2016. Shedding light on the 'dark side' of phylogenetic
562 comparative methods. Methods Ecol. Evol. 7, 693–699. https://doi.org/10.1111/2041-
563 210X.12533

564 Danforth, B., Minckley, R., Neff, J., 2019. The solitary bees: biology, evolution, conservation.
565 Princeton University Press.

566 Danforth, B.N., Eardley, C., Packer, L., Walker, K., Pauly, A., Randrianambinintsoa, F.J., 2008.
567 Phylogeny of Halictidae with an emphasis on endemic African Halictinae. Apidologie 39, 86–
568 101. https://doi.org/10.1051/apido:2008002

569 Davison, P.J., Field, J., 2018. Limited social plasticity in the socially polymorphic sweat bee
570 Lasioglossum calceatum. Behav. Ecol. Sociobiol. 72. https://doi.org/10.1007/s00265-018-2475-
571 9

572 Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the
573 multispecies coalescent. Trends Ecol. Evol. 24, 332–340.
574 https://doi.org/10.1016/j.tree.2009.01.009

575 Driskell, A.C., Ané, C., Burleig, J.G., McMahon, M.M., O'Meara, B.C., Sanderson, M.J., 2004. Prospects
576 for building the tree of life from large sequence databases. Science (80-.). 306, 1172–1174.
577 https://doi.org/10.1126/science.1102036

578 Eardley, C.D., Gikungu, M., Schwarz, M.P., 2009. Bee conservation in Sub-Saharan Africa and
579 Madagascar: Diversity, status and threats. Apidologie 40, 355–366.
580 https://doi.org/10.1051/apido/2009016

581 Fisher, A.A., Hassler, G.W., Ji, X., Baele, G., Suchard, M.A., Lemey, P., 2022. Scalable Bayesian
582 phylogenetics. Philos. Trans. R. Soc. B Biol. Sci. 377. https://doi.org/10.1098/rstb.2021.0242

583 Freitas, B.M., Imperatriz-Fonseca, V.L., Medina, L.M., Kleinert, A.D.M.P., Galetto, L., Nates-Parra, G.,
584 Javier, J., 2009. Diversity, threats and conservation of native bees in the Neotropics. Apidologie
585 40, 332–346. https://doi.org/10.1051/apido/2009012

586 Freitas, F. V., Branstetter, M.G., Casali, D.M., Aguiar, A.J.C., Griswold, T., Almeida, E.A.B., 2022.
587 Phylogenomic dating and Bayesian biogeography illuminate an antitropical pattern for eucerine
588 bees. J. Biogeogr. 49, 1034–1047. https://doi.org/10.1111/jbi.14359

589 Freyman, W.A., 2015. SUMAC: Constructing phylogenetic supermatrices and assessing partially
590 decisive taxon coverage. Evol. Bioinforma. 11, 263–266. https://doi.org/10.4137/EBO.S35384

591 Garamszegi, L.Z., Møller, A.P., 2011. Nonrandom variation in within-species sample size and missing
592 data in phylogenetic comparative studies. Syst. Biol. 60, 876–880.
593 https://doi.org/10.1093/sysbio/syr060

594 Gonzalez, V.H., Griswold, T., Praz, C.J., Danforth, B.N., 2012. Phylogeny of the bee family
595 Megachilidae (Hymenoptera: Apoidea) based on adult morphology. Syst. Entomol. 37, 261–
596 286. https://doi.org/10.1111/j.1365-3113.2012.00620.x

597 Groom, S.V.C., Schwarz, M.P., 2011. Bees in the southwest pacific: Origins, diversity and
598 conservation. Apidologie 42, 759–770. https://doi.org/10.1007/s13592-011-0079-8

599 Hahn, M.W., Nakhleh, L., 2016. Irrational exuberance for resolved species trees. Evolution (N. Y). 70,
600 7–17. https://doi.org/10.1111/evo.12832

601 Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program
602 for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41, 95–98.

603 Harvey, P.H., Pagel, M.D., 1991. The Comparative Method in Evolutionary Biology. Oxford university
604 press Oxford.

605 Hedtke, S.M., Patiny, S., Danforth, B.N., 2013. The bee tree of life: A supermatrix approach to apoid
606 phylogeny and biogeography. BMC Evol. Biol. 13. https://doi.org/10.1186/1471-2148-13-138

607 Hoang, D.T., Chernomor, O., Von Haeseler, A., Minh, B.Q., Vinh, L.S., 2018. UFBoot2: Improving the
608 ultrafast bootstrap approximation. Mol. Biol. Evol. 35, 518–522.
609 https://doi.org/10.1093/molbev/msx281

610 Hosner, P.A., Zhao, M., Kimball, R.T., Braun, E.L., Burleigh, J.G., 2022. Updating splits, lumps, and
611 shuffles: Reconciling GenBank names with standardized avian taxonomies. Ornithology 139, 1–
612 15. https://doi.org/10.1093/ornithology/ukac045

613 Hugall, A.F., Foster, R., Lee, M.S.Y., 2007. Calibration choice, rate smoothing, and the pattern of
614 tetrapod diversification according to the long nuclear gene RAG-1. Syst. Biol. 56, 543–563.
615 https://doi.org/10.1080/10635150701477825

616 Hugall, A.F., Stuart-Fox, D., 2012. Accelerated speciation in colour-polymorphic birds. Nature 485,
617 631–634. https://doi.org/10.1038/nature11050

618 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., Bui Quang, M., Wong, T.K.F., von Haeseler, A., Jermiin, L.S., 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. 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., Laurent, R.A.S., 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., Childers, R.A.R., 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. 1–11. 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. https://doi.org/10.3390/d11070109

Kuhn, T.S., Mooers, A., Thomas, G.H., 2011. A simple polytomy resolver for dated phylogenies.

651 Methods Ecol. Evol. 2, 427–436. https://doi.org/10.1111/j.2041-210X.2011.00103.x

652 Litman, J.R., Danforth, B.N., Eardley, C.D., Praz, C.J., 2011. Why do leafcutter bees cut leaves? new
653 insights into the early evolution of bees. Proc. R. Soc. B Biol. Sci. 278, 3593–3600.
654 https://doi.org/10.1098/rspb.2011.0365

655 Martins, A.C., Melo, G.A.R., Renner, S.S., 2014. The corbiculate bees arose from New World oil-
656 collecting bees: Implications for the origin of pollen baskets. Mol. Phylogenet. Evol. 80, 88–94.
657 https://doi.org/10.1016/j.ympev.2014.07.003

658 McCraney, W.T., Thacker, C.E., Alfaro, M.E., 2020. Supermatrix phylogeny resolves goby lineages and
659 reveals unstable root of Gobiaria. Mol. Phylogenet. Evol. 151, 106862.
660 https://doi.org/10.1016/j.ympev.2020.106862

661 McFrederick, Q.S., Wcislo, W.T., Hout, M.C., Mueller, U.G., 2014. Host species and developmental
662 stage, but not host social structure, affects bacterial community structure in socially
663 polymorphic bees. FEMS Microbiol. Ecol. 88, 398–406. https://doi.org/10.1111/1574-
664 6941.12302

665 McGowen, M.R., Spaulding, M., Gatesy, J., 2009. Divergence date estimation and a comprehensive
666 molecular tree of extant cetaceans. Mol. Phylogenet. Evol. 53, 891–906.
667 https://doi.org/10.1016/j.ympev.2009.08.018

668 Melin, A., Colville, J.F., 2019. A review of 250 years of South African bee taxonomy and exploration
669 (Hymenoptera: Apoidea: Anthophila). Trans. R. Soc. South Africa 74, 86–96.
670 https://doi.org/10.1080/0035919X.2019.1572670

671 Mello, B., Schrago, C.G., 2014. Assignment of Calibration Information to Deeper Phylogenetic Nodes
672 is More Effective in Obtaining Precise and Accurate Divergence Time Estimates. Evol.
673 Bioinforma. 10, 79–85. https://doi.org/10.4137/EBO.S13908

674 Michener, C.D., 2007. The Bees of the World. Johns Hopkins University Press, Baltimore.

675 Michez, D., Patiny, S., Danforth, B.N., 2009. Phylogeny of the bee family Melittidae (Hymenoptera:
676 Anthophila) based on combined molecular and morphological data. Syst. Entomol. 34, 574–
677 597. https://doi.org/10.1111/j.1365-3113.2009.00479.x

678 Microsoft Corporation, 2021. Microsoft Excel.

679 Miller, M.A., Pfeiffer, W., Schwartz, T., 2010. Creating the CIPRES Science Gateway for inference of
680 large phylogenetic trees. 2010 Gatew. Comput. Environ. Work. GCE 2010.

https://doi.org/10.1109/GCE.2010.5676129

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

Mora, C., Tittensor, D.P., Adl, S., Simpson, A.G.B., Worm, B., 2011. How many species are there on earth and in the ocean? PLoS Biol. 9, 1–8. https://doi.org/10.1371/journal.pbio.1001127

Mouquet, N., Devictor, V., Meynard, C.N., Munoz, F., Bersier, L.F., Chave, J., Couteron, P., Dalecky, A., Fontaine, C., Gravel, D., Hardy, O.J., Jabot, F., Lavergne, S., Leibold, M., Mouillot, D., Münkemüller, T., Pavoine, S., Prinzing, A., Rodrigues, A.S.L., Rohr, R.P., Thébault, E., Thuiller, W., 2012. Ecophylogenetics: Advances and perspectives. Biol. Rev. 87, 769–785. https://doi.org/10.1111/j.1469-185X.2012.00224.x

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

Near, T.J., Meylan, P.A., Shaffer, H.B., 2005. Assessing concordance of fossil calibration points in molecular clock studies: An example using turtles. Am. Nat. 165, 137–146. https://doi.org/10.1086/427734

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. Soc. 1–16. https://doi.org/10.5061/dryad.nk98sf7x7

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

Orr, M.C., Hughes, A.C., Chesters, D., Pickering, J., Zhu, C.-D., Ascher, J.S., 2020. Global Patterns and Drivers of Bee Distribution. Curr. Biol. 31, 451–458. https://doi.org/10.1016/j.cub.2020.10.053

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., 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

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

Pisanty, G., Richter, R., Martin, T., Dettman, J., Cardinal, S., 2022. 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. Insectes Soc. 47, 263–270.

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 (N. Y). 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.

743 Phylogenetic uncertainty revisited: Implications for ecological analyses. Evolution (N. Y). 69,
744 1301–1312. https://doi.org/10.1111/evo.12644

745 Rees, J.A., Cranston, K., 2017. Automated assembly of a reference taxonomy for phylogenetic data
746 synthesis. Biodivers. Data J. 5. https://doi.org/10.3897/BDJ.5.e12581

747 Rehan, S., Schwarz, M., 2015. A few steps forward and no steps back: Long-distance dispersal
748 patterns in small carpenter bees suggest major barriers to back-dispersal. J. Biogeogr. 42, 485–
749 494. https://doi.org/10.1111/jbi.12439

750 Rehan, S.M., Chapman, T.W., Craigie, A.I., Richards, M.H., Cooper, S.J.B., Schwarz, M.P., 2010.
751 Molecular phylogeny of the small carpenter bees (Hymenoptera: Apidae: Ceratinini) indicates
752 early and rapid global dispersal. Mol. Phylogenet. Evol. 55, 1042–1054.
753 https://doi.org/10.1016/j.ympev.2010.01.011

754 Richards, M.H., Von Wettberg, E.J., Rutgers, A.C., 2003. A novel social polymorphism in a primitively
755 eusocial bee. Proc. Natl. Acad. Sci. U. S. A. 100, 7175–7180.
756 https://doi.org/10.1073/pnas.1030738100

757 Rosa, B.B., Melo, G.A.R., 2021. Apoid wasps (Hymenoptera: Apoidea) from mid-Cretaceous amber of
758 northern Myanmar. Cretac. Res. 122, 104770. https://doi.org/10.1016/j.cretres.2021.104770

759 Roure, B., Baurain, D., Philippe, H., 2013. Impact of missing data on phylogenies inferred from
760 empirical phylogenomic data sets. Mol. Biol. Evol. 30, 197–214.
761 https://doi.org/10.1093/molbev/mss208

762 Sanderson, M.J., 2002. Estimating absolute rates of molecular evolution and divergence times: A
763 penalized likelihood approach. Mol. Biol. Evol. 19, 101–109.
764 https://doi.org/10.1093/oxfordjournals.molbev.a003974

765 Sanderson, M.J., McMahon, M.M., Steel, M., 2010. Phylogenomics with incomplete taxon coverage:
766 The limits to inference. BMC Evol. Biol. 10. https://doi.org/10.1186/1471-2148-10-155

767 Sanderson, M.J., Purvis, A., Henze, C., 1998. Phylogenetic supertrees: Assembling the trees of life.
768 Trends Ecol. Evol. 13, 105–109. https://doi.org/10.1016/S0169-5347(97)01242-1

769 Sann, M., Meusemann, K., Niehuis, O., Escalona, H.E., Mokrousov, M., Ohl, M., Pauli, T., Schmid-
770 Egger, C., 2021. Reanalysis of the apoid wasp phylogeny with additional taxa and sequence
771 data confirms the placement of Ammoplanidae as sister to bees. Syst. Entomol. 46, 558–569.
772 https://doi.org/10.1111/syen.12475

773 Sann, M., Niehuis, O., Peters, R.S., Mayer, C., Kozlov, A., Podsiadlowski, L., Bank, S., Meusemann, K.,
774 Misof, B., Bleidorn, C., Ohl, M., 2018. Phylogenomic analysis of Apoidea sheds new light on the
775 sister group of bees. BMC Evol. Biol. 18, 1–15. https://doi.org/10.1186/s12862-018-1155-8

776 Schwarz, M., Fuller, S., Tierney, S., Cooper, S.J., 2006. Molecular phylogenetics of the exoneurine
777 allodapine bees reveal an ancient and puzzling dispersal from Africa to Australia. Syst. Biol. 55,
778 31–45. https://doi.org/10.1080/10635150500431148

779 Shen, X.X., Li, Y., Hittinger, C.T., Chen, X. xin, Rokas, A., 2020. An investigation of irreproducibility in
780 maximum likelihood phylogenetic inference. Nat. Commun. 11, 1–14.
781 https://doi.org/10.1038/s41467-020-20005-6

782 Sless, T.J.L., Branstetter, M.G., Gillung, J.P., Krichilsky, E.A., Tobin, K.B., Straka, J., Rozen, J.G., Freitas,
783 F. V., Martins, A.C., Bossert, S., Searle, J.B., Danforth, B.N., 2022. Phylogenetic relationships and
784 the evolution of host preferences in the largest clade of brood parasitic bees (Apidae:
785 Nomadinae). Mol. Phylogenet. Evol. 166, 107326.
786 https://doi.org/10.1016/j.ympev.2021.107326

787 Smirnov, V., Warnow, T., 2021. Phylogeny Estimation Given Sequence Length Heterogeneity. Syst.
788 Biol. 70, 268–282. https://doi.org/10.1093/sysbio/syaa058

789 Smith, M.R., 2022. Using Information Theory to Detect Rogue Taxa and Improve Consensus Trees.
790 Syst. Biol. 71, 1088–1094. https://doi.org/10.1093/sysbio/syab099

791 Smith, S.A., Brown, J.W., 2018. Constructing a broadly inclusive seed plant phylogeny. Am. J. Bot.
792 105, 302–314. https://doi.org/10.1002/ajb2.1019

793 Smith, S.A., O'Meara, B.C., 2012. treePL: Divergence time estimation using penalized likelihood for
794 large phylogenies. Bioinformatics 28, 2689–2690.
795 https://doi.org/10.1093/bioinformatics/bts492

796 Soltis, P.S., Soltis, D.E., Wolf, P.G., Nickrent, D.L., Chaw, S.M., Chapman, R.L., 1999. The phylogeny of
797 land plants inferred from 18S rDNA sequences: Pushing the limits of rDNA signal? Mol. Biol.
798 Evol. 16, 1774–1784. https://doi.org/10.1093/oxfordjournals.molbev.a026089

799 Stamatakis, A., 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large
800 phylogenies. Bioinformatics 30, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

801 Thomas, G.H., Hartmann, K., Jetz, W., Joy, J.B., Mimoto, A., Mooers, A.O., 2013. PASTIS: An R
802 package to facilitate phylogenetic assembly with soft taxonomic inferences. Methods Ecol.
803 Evol. 4, 1011–1017. https://doi.org/10.1111/2041-210X.12117

Trunz, V., Packer, L., Vieu, J., Arrigo, N., Praz, C.J., 2016. Comprehensive phylogeny, biogeography and new classification of the diverse bee tribe Megachilini: Can we use DNA barcodes in phylogenies of large genera? Mol. Phylogenet. Evol. 103, 245–259. https://doi.org/10.1016/j.ympev.2016.07.004

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 Biology. 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.

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

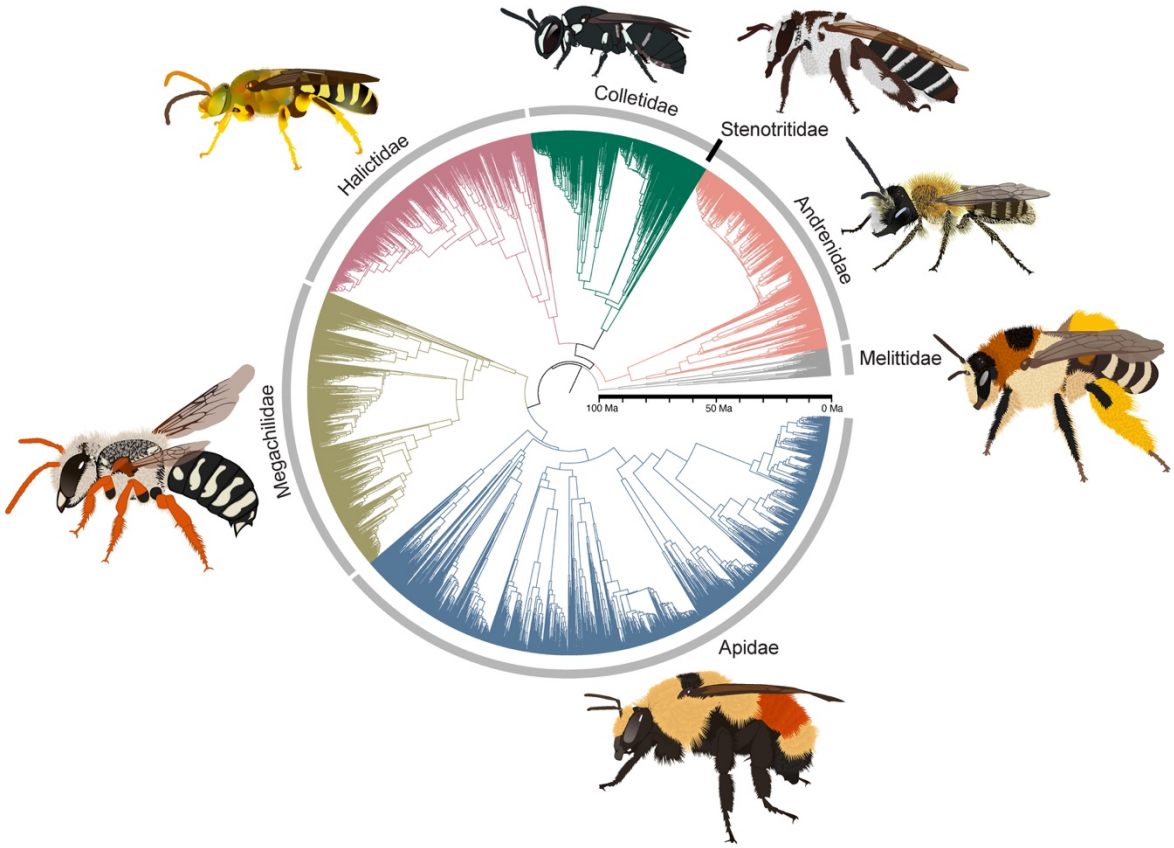

Fig. 1. Supermatrix phylogeny for Anthophila built with public and published sequence data. The full tree is available for download at beetreeoflife.org and includes 4651 species in 427 genera, including the species illustrated in the figure: *Agapostemon texanus* (Halictidae), *Andrena fulva* (Andrenidae), *Anthidium chilense* (Megachilidae), *Bombus huntii* (Apidae), *Ctenocolletes nigricans* (Stenotritidae), *Dasypoda hirtipes* (Melittidae), and *Hylaeus affinis* (Colletidae). Family colour coding used throughout.

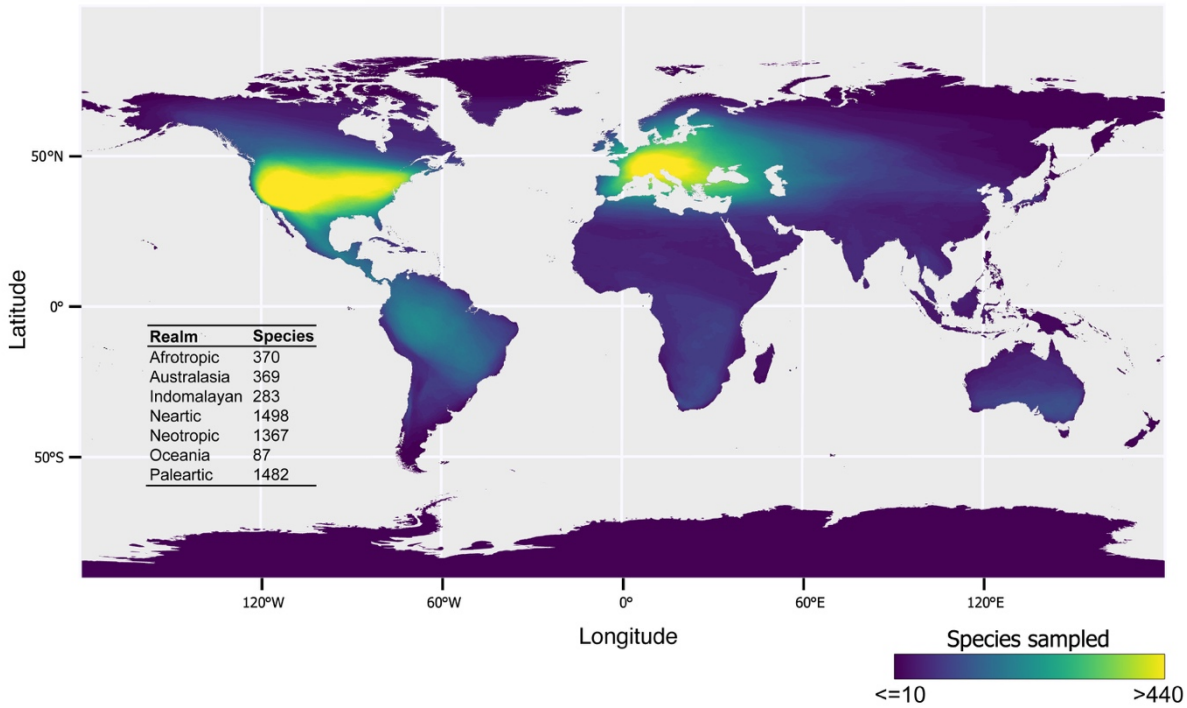

Fig. 2. Distribution of bees with available molecular data that was used to build the supermatrix phylogeny.

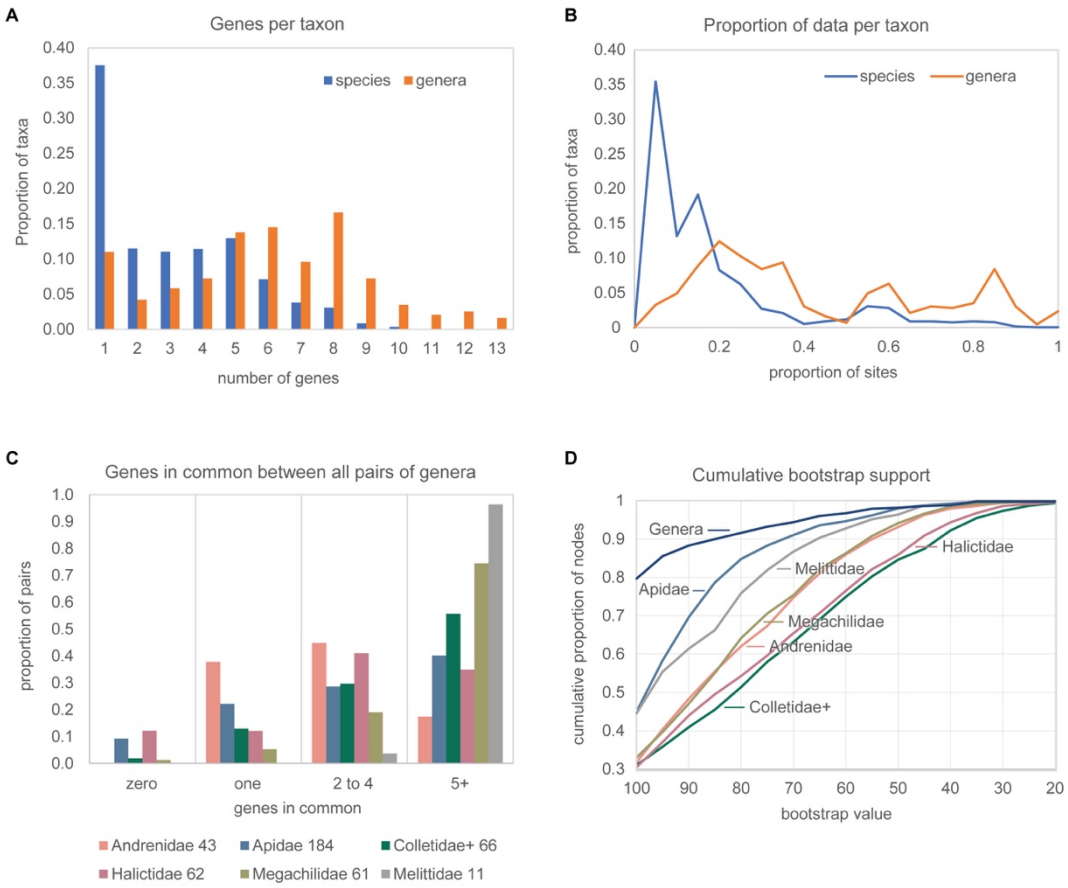

Fig. 3. Graphical summaries of data completeness and tree support. **A.** Number of genetic data elements ('genes') per species in the 4651 species tree and per genus in the 427 genera representative tree. **B.** Proportion of sites with data in the supermatrix not including the phylogenomic stub. **C.** Number of genes 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.