Article title

SUPERSMART: Ecology and Evolution in the Era of Big Data

Authors' names

Alexandre Antonelli[1,2,*], Fabien L. Condamine[1], Hannes Hettling[3], Karin Nilsson[1,3], R. Henrik Nilsson[1], Bengt Oxelman[1], Michael J. Sanderson[4], Hervé Sauquet[5], Ruud Scharn[1], Daniele Silvestro[1,6], Mats Töpel[1,7], Rutger A. Vos[3,*]

Authors' affiliations

[1] *University of Gothenburg, Department of Biological and Environmental Sciences, Box 461, SE-405 30 Göteborg, Sweden*

[2] *Gothenburg Botanical Garden, Carl Skottsbergs gata 22A, SE-41319, Göteborg, Sweden*

[3] *Naturalis Biodiversity Center, Darwinweg 4, 2333 CR Leiden, the Netherlands*

[4] *University of Arizona, Department of Ecology and Evolutionary Biology, 1041 E. Lowell, Tucson, USA*

[5] *Université Paris-Sud, Laboratoire Écologie, Systématique, Évolution, CNRS UMR 8079, 91405 Orsay, France*

[6] *University of Lausanne, Department of Ecology and Evolution, 1015 Lausanne, Switzerland*

[7] *Swedish Bioinformatics Infrastructure for Life Sciences, Sweden*

Short running title

The SUPERSMART approach

Keywords: bioinformatics, biogeography, dated tree of life, diversification, macroecology, open science, phylogenetics.

Authors' contributions

A.A. and R.A.V. coordinated the study and led the writing with contributions from all authors. A.A., R.H.N., B.O., D.S. and M.T. initiated the project; R.A.V., K.N. and H.H. wrote the code; H.S. and F.L.C. developed the fossil calibration database; R.S. and H.H. performed the Gentianales analysis; M.J.S. and R.A.V. further developed the PhyLoTa browser.

Corresponding authors (*)

Alexandre Antonelli: University of Gothenburg, Department of Biological and Environmental Sciences, Box 461, SE-405 30 Göteborg, Sweden

& Gothenburg Botanical Garden, Carl Skottsbergs gata 22A, SE-41319, Göteborg, Sweden

alexandre.antonelli@bioenv.gu.se

Rutger Vos: Naturalis Biodiversity Center, Darwinweg 4, 2333 CR Leiden, the Netherlands

rutger.vos@naturalis.nl

ABSTRACT

Rapidly growing biological data volumes – including molecular sequences, species traits, geographic occurrences, specimen collections, and fossil records – hold an unprecedented, yet largely unexplored potential to reveal how ecological and evolutionary processes generate and maintain biodiversity. Most biodiversity studies integrating ecological data and evolutionary history use an idiosyncratic step-by-step approach for the reconstruction of time-calibrated phylogenies in light of ecological and evolutionary scenarios. Here we introduce a conceptual framework, termed SUPERSMART (Self-Updating Platform for Estimating Rates of Speciation and Migration, Ages, and Relationships of Taxa), and provide a proof of concept for dealing with the moving targets of biodiversity research. This framework reconstructs dated phylogenies based on the assembly of molecular datasets and collects pertinent data on ecology, distribution, and fossils of the focal clade. The data handled for each step are continuously updated as databases accumulate new records. We exemplify the practice of our method by presenting comprehensive phylogenetic and dating analyses for the orders Primates and the Gentianales. We believe that this emerging framework will provide an invaluable tool for a wide range of hypothesis-driven research questions in ecology and evolution.

INTRODUCTION

"We are drowning in information, while starving for wisdom"

E. O. Wilson, *Consilience: The unity of knowledge* (1999)

In biodiversity studies, many important theories have been built through the accumulation of data (e.g. ecological, genetic, and geographic) in order to unveil most of today's well-known patterns of biodiversity, ranging from the latitudinal diversity gradient, through the species-abundance distribution, to descent with modification. Nowadays, the amount of data available for studying many aspects of biodiversity is tantalizing (see Fig. 1 for key examples). For instance, DNA records of the International Nucleotide Sequence Database Collaboration (INSDC) including GenBank, ENA, and DDBJ (Nakamura *et al*. 2013) grow exponentially, doubling in number every 10 months and currently comprising over 150 million sequences from more than 300,000 species. Similarly, there are over 445 million observation records available through the Global Biodiversity Information Facility (http://www.gbif.org) and over 1.1 million taxonomic records of fossils in the Paleobiology Database (http://fossilworks.org/). In addition, ecological data are increasingly compiled and stored in archives or repositories for entire clades, e.g. mammals in PanTHERIA (Jones *et al*. 2009), birds in eBird (http://ebird.org), and plants in TRY (http://try-db.org). Despite important caveats concerning the uneven geographic, temporal, and taxonomic representation in these databases, as well as varying levels of quality and annotation (Bidartondo *et al*. 2008), it is clear that these data hold a tremendous – yet largely unexplored – scientific potential and explanatory power. However, it is less clear how these large data collections are

best integrated in biodiversity research, and what effect they may have on understanding ecological and evolutionary processes that have shaped diversity patterns through time and space – from estimates of species delimitations, species relationships, divergence times, and diversification rates, to historical biogeography and macroecology. Nonetheless, if we want to start assembling the big picture to reveal how ecological and evolutionary processes generate and maintain biodiversity (Zanne *et al*. 2014), we need approaches that integrate ecological data and evolutionary history in a user-friendly framework to study biodiversity at various temporal, spatial and taxonomic scales (Chave 2013).

In this paper we review the current challenges in evolutionary research for reaching this goal and highlight the prospects of biodiversity workflows. We also identify and discuss solutions for inherent concerns in the quality and completeness of the data we are handling. We then present a new conceptual framework, the *Self-Updating Platform for Estimating Rates of Speciation and Migration, Ages and Relationships of Taxa* (SUPERSMART), which allows researchers to use all publicly available genetic, genomic, ecological, and geographic data available, in addition to their own data. We illustrate this approach with two empirical examples, the Primates (including humans, apes, monkeys and prosimians) and the plant order Gentianales (including the coffee family Rubiaceae and the dogbane family Apocynaceae, among others).

OUTSTANDING CHALLENGES IN ECO-EVOLUTIONARY RESEARCH

The Tree of Life. Assembling a complete species-level tree (or network) of life constitutes an overarching goal in biology. A major obstacle is that genetic sampling of species is taxonomically and geographically highly biased (Gotelli & Colwell 2001). Wealthy but species-poor countries in the Northern Hemisphere are generally better sampled than tropical, species-rich developing regions in Latin America, Africa, and Southeast Asia, although this situation may be changing with developing countries such as Brazil investing heavily in molecular projects in conjunction with scientific education. A second hurdle is the fact that scientists have used different sets of genes and genetic markers for different taxa, both for intrinsic reasons (e.g. markers differ in information content among taxa, ease of sequencing, and quality of source material) and because of a lack of consensus on which markers to use for addressing similar phylogenetic problems.

Two main approaches have been developed to take advantage of the sequencing and phylogenetic efforts done so far, both of which have the capacity to handle very large numbers of terminal taxa: i) *supertrees*, which involve the fusion of separate trees that should have at least some degree of taxonomic overlap, using parsimony, maximum likelihood, or Bayesian approaches (e.g., Nguyen *et al*. 2012, and references therein); and ii) *supermatrices*, which are datasets containing sets of markers that may partly overlap, such that not all taxa are covered by all markers (de Queiroz & Gatesy 2007). A schematic illustration depicting the rationale and differences between supertrees and supermatrices is provided in Fig. 2. Both approaches present

particular advantages as well as limitations (von Haeseler 2012), and alternatives are starting to develop (Smith *et al*. 2013).

Supertrees have so far been the only solution to produce a single, near-complete phylogenetic tree comprising all organisms in a clade. They can be built even when there is no genetic overlap among the subtrees they comprise, their inference is usually fast, and their mathematical properties well studied; factors which jointly have made supertrees (or variations thereof) the preferred choice for current synthetic projects such as the Open Tree of Life (http://www.opentreeoflife.org). Criticism against supertrees includes the difficulties – both practical and theoretical – in combining tree topologies and computing clade support from trees derived from different sources of data, and often also different methods and assumptions. Supertree approaches have also been shown through simulations to be less accurate than supermatrices in recovering correct topologies (Kupczok *et al*. 2010). Another factor that has contributed to defavoring supertrees is the realisation that just a small fraction of phylogenetic trees published can be retrieved through open data repositories or direct requests to authors of phylogenetic papers (Drew 2013; Stoltzfus *et al*. 2013). In a recent study, it was only possible to obtain trees from about 17% of 7500+ phylogenetic articles from the last 12 years – a figure that also includes trees with poor or inconsistent underlying data such as incomplete taxon names, lack of information on which characters were used in the analyses, and missing settings and input files for phylogenetic analyses (Drew *et al*. 2013).

Supermatrix approaches allow the estimation of large trees under a single analysis, relying directly on the underlying data (molecular and/or morphological) rather than on tree topologies. Analyses of empirical and simulated data suggest that initial concerns with missing data may have been largely unsubstantiated. Even a relatively small set of informative characters, such as a single gene or genetic marker scored across all taxa, may potentially provide the backbone of a phylogeny and allow more rapidly evolving markers to resolve terminal relationships (Wiens 1998, 2006; Kupczok *et al*. 2010). The total number of terminals in a supermatrix has long been a limiting factor for tree reconstruction, but increasingly larger phylogenetic trees can now be inferred using both parsimony and maximum likelihood methods (Sanderson 2008; Goloboff *et al*. 2009; Stamatakis *et al*. 2010). In addition, existing phylogenetic algorithms are constantly being optimized (Stamatakis *et al*. 2012) and new ones introduced (Price *et al*. 2010).

A major drawback with supermatrices spanning large taxonomic units and evolutionary times is homology assessment during the alignment of highly divergent or saturated sequences. Although this issue is often ignored or not formally dealt with (e.g. only through visual inspection of sequences and manual exclusion of apparent outliers), automated methods have been developed to detect rogue taxa (Aberer *et al*. 2013), sequence saturation, and perform profile alignment of very large supermatrices (Smith *et al*. 2009). In addition, current initiatives now aim at standardizing the identification and annotation of orthologous genes and their phenotypes, which will certainly facilitate the generation of aligned supermatrices containing only orthologs (e.g. http://inparanoid.sbc.su.se and http://www.phenotypercn.org). Although

these efforts address some of the difficulties in assembling input datasets, the simple application of supermatrix approaches to recover both deep relationships among higher taxa as well as more recent divergences in one analysis may not always succeed. This is due to limitations in the scalability of current state-of-the-art phylogenetic software for recovering species trees from gene trees (Edwards 2009). Supermatrix analyses therefore typically resort to less sophisticated, but more scalable techniques. A serious shortcoming of both supertree and supermatrix methods is that they typically assume that all data partitions are evolving according to the same tree, thus failing to account for processes such as incomplete lineage sorting, hybridisation, and gene duplications/losses (Whidden *et al*. 2014).

The Chronogram of Life. Estimating divergence times among all species poses similar as well as novel challenges as compared to phylogenetic inference. In the absence of full sequence coverage, dating supertrees requires hybrid approaches that include both fossil-calibrated sequence data (Vos & Mooers 2004) as well as, for nodes lacking sequence coverage, the application of expected waiting times between speciation events based on models of clade growth (Gernhard *et al*. 2006). Edge length estimation may be just as challenging for large, gappy supermatrices.

There are several methods and software suites available for molecular dating, each with its own set of assumptions, advantages, and potential caveats. Even if the choice of dating method may lead to substantial differences in age estimations (Linder *et al*. 2005; Gustafsson *et al*. 2010), node calibration (Sauquet *et al*. 2012) and prior distributions on ages (Ho & Phillips

2009) are often the most crucial steps. Time calibration can be done *directly* or *indirectly* (also termed primary and secondary calibration, respectively).

Direct calibration enforces age constraints on specific clades of the phylogeny, most often through the use of fossils but sometimes through geological events (such as the age of an island or a land bridge). The location of calibration points in the tree appears to determine the extent to which variance propagates upwards from the tip to the root (for recent calibration points) or is constrained (for older calibration points; Vos & Mooers 2004). While early dating methods usually required those ages to be fixed, or to be "hard" minimum or maximum age constraints (Sanderson 2002, 2003), it is now possible to model uncertainties in the timing information as "soft" prior probability distributions (Drummond *et al*. 2006; Ho & Phillips 2009).

Indirect calibrations typically rely either on applying nucleotide substitution rates derived in other studies of closely related taxa, or on using the age of a lineage split estimated in a previous dating analysis. Indirect calibration is practical when there are no fossils or other direct age constraints available for the focal group. Although this may result in increased uncertainties in estimated ages, it is possible to transfer posterior age distributions from one Bayesian analysis as priors to another. In addition, calibration may be improved by implementing it on two or more nodes of the phylogeny rather than a single one (Sauquet *et al*. 2012).

Considering the many methodological options available and the complexity of working with imperfect empirical data, it is not surprising that studies employing molecular dating

analyses show a wide spectrum of variation. This includes the various uses of available software; the varying quality and reliability of the fossil record (in terms of phylogenetic placement, absolute age, and proximity to the true timing of speciation of the taxon they represent), and the reliability of the molecular data supporting the chronograms. Although the evaluation of dating methodologies and assumptions will certainly continue for the foreseeable future, based on these considerations few would contest that *estimated ages from different studies are not directly comparable*. This crucial realisation suggests that dated phylogenies cannot be reliably 'pasted together' in a similar way as traditional supertrees. Moreover, this cautions against the increasingly widespread use of dated phylogenies of various sources in meta-analyses, despite their potential as a powerful way of studying macroevolutionary processes, including the historical assembly of biomes (Crisp *et al*. 2009; Hoorn *et al*. 2010), dispersal across biotic barriers (Cody *et al*. 2010), or correlations between lineage age and diversity (Rabosky *et al*. 2012). It remains to be assessed to what extent meta-analyses of published chronograms are able to identify significant signal amid the expected background 'noise' of erroneous dating estimates.

Dealing with conflict. There is an ongoing paradigm shift in evolutionary biology, where single gene trees, consensus, and concatenation approaches are replaced by integrative species tree thinking (Edwards 2009). Although the distinction between gene and species trees is not new (Pamilo & Nei 1988), many species phylogenies published to date are based on single gene trees or trees derived from the concatenation of two or more alignments from different linkage groups (e.g. as in the supermatrix approach). However, models and methods have been developed to account for the fact that gene trees can differ in topology and branch lengths due to population

genetics processes that can be modelled with the multispecies coalescent (Rannala & Yang 2003). Ignoring conflict may influence not only tree topologies but also branch lengths, measured in time. Splits in gene trees overestimate species divergence times, and the size of this bias is determined by ancestral effective population sizes and branch lengths of the species tree. One accurate Bayesian implementation where species trees can be inferred directly from aligned sequence data is the multi-species, multi-marker approach implemented in *BEAST (Heled & Drummond 2010). For larger taxon sets, however, simpler approaches where gene trees are used as input data may be required (Bayzid & Warnow 2012).

DATA QUALITY CONCERNS

As the volume of data grows, it is not always possible to verify the quality of biological data manually, such that much of the quality control has to be handed over to algorithms. However, some aspects of biological research are not readily amenable to algorithmic quality control, such as the concepts of species and species delimitation (Chesters & Vogler 2013), although there are new promising developments in this field (Fujita *et al*. 2012). Thus, data growth in itself is not tantamount to an immediate corresponding growth in knowledge.

Genetic data. Fungi provide a compelling example: more than half of all ~350,000 fungal ribosomal ITS sequences – the formal fungal barcode – are annotated to various uninformative levels such as "Uncultured fungus" (Bengtsson-Palme *et al*. 2013). Of the fungal ITS sequences that do have a species name, more than 10% carry an incorrect name due to misidentification,

contamination, or technical complications; similar estimates have been produced for other groups of organisms (Valkiunas *et al*. 2008; de Mendonça *et al*. 2011). The situation with annotating sequences to gene level is often not much better: the ribosomal small subunit (16S) forms the standard marker in prokaryote molecular ecology, yet anyone who seeks to manually download 16S sequences from INSDC will have to explore a near-endless array of orthographic and conceptual variations such as "16 S", "16S", "17S", "SSU", "ribosomal small subunit", and "ribosomal small sub-unit". Many sequence authors are slow to update their records with recent and correct information, forming an additional obstacle for anyone using these databases (Hyde *et al*. 2013).

Species distributions. Data quality concerns are not unique to repositories of molecular sequences; they are at least as serious for other core biological data such as species occurrences. The Global Biodiversity Information Facility (GBIF; http://www.gbif.org), the main portal for accessing locality data worldwide, aggregates numerous databases of natural history collections and species observations, of which about 85% (c. 381 million) are currently geo-referenced (Fig. 1B). At least five potential problems prevent the widespread use of this vast amount of data: *incorrectly geo-referenced records; lumping of native and non-native occurrences; erroneous taxonomic identifications; name synonymisation issues*; and *sampling biases*. Some of these issues are easier to tackle than others. Terrestrial plants appearing in the middle of the ocean due to a switch in latitude/longitude (or due to the use of different coordinate projection systems) are easier to spot than animals in a zoo or cultivated plants in a botanical garden outside of their native ranges (Yesson *et al*. 2007). The data cleaning workflow implemented in the Biology

Virtual e-Laboratory platform (http://www.biovel.eu) is one example of how bioinformatics may aid researchers to visualise and clean species locality data. It is important to realise, however, that the level of spatial resolution required by different scientific questions varies greatly. For instance, whereas very precise and verified records are required by ecological niche modelling of species distributions (Duputié *et al*. 2014), biogeographic analyses at the continental level would be much less sensitive to imprecise georeferencing. With more than one billion biological specimens currently stored in the world's natural history collections, this means that even a rough and automated tagging of those collections may greatly contribute to biodiversity research (Guralnick *et al*. 2006) and hopefully ameliorate the serious spatial, taxonomic, and temporal biases in currently available species occurrence data (Boakes *et al*. 2010).

Fossils. The deposition of fossil information in open data repositories has not been as successful as for genetic data, rendering palaeobiological databases such as the Paleobiology database (http://fossilworks.org; Fig. 1C), New and Old Worlds (NOW, http://www.helsinki.fi/science/now/index.html), and the Neotoma Paleoecology Database (http://www.neotomadb.org) apparently even more taxonomically and geographically biased than INSDC and GBIF. This is partly due to the fact that several paleontological journals still lack strong data deposition policies (or fall short of enforcing them), but also because some important works have not yet been digitized, e.g. the Cenozoic Mammals of Africa (Werdelin & Sanders 2010). One must also keep in mind that data may have been entered for different reasons. While some fossil records in the Paleobiology Database will stem from meticulously undertaken taxonomic studies, e.g. using scanning electron microscopy techniques for the identification of

fossil pollen, others derive from ecological quantitative studies where identification was done under simpler, less precise methods using different species concepts or identified only at high taxonomic levels such as genus, family, or order (Johnsrud *et al*. 2013). In most instances, such records were never entered with the intention of being used as calibration points for molecular dating analyses, a procedure that requires careful examination of the phylogenetic relationships of fossil taxa, in particular the distinction between stem *vs*. crown-group lineages (e.g., Sauquet *et al*. 2012).

'Large and dirty' or 'small and clean'? Given the data quality concerns outlined above, managers of current and new biological databases face a hard decision: to either enable an easy but poorly controlled input of new data, or to enforce control measures that maintain a higher quality level but may lead to reduced data influx. As an example, Ksepka et al. (2011) proposed a new fossil data resource – the Fossil Calibration Database – specifically aiming at selecting fossils suitable for molecular dating analyses, where fossils included need to comply with a set of five pre-defined criteria and are subjected to peer-review in the form of manuscripts sent to the *Paleontologia Electronica* (PE) journal (Parham *et al*. 2012a). The Fossil Calibration Database will not only contain information on adequate fossils for calibration but also their phylogenetic placement and age. Other examples of open databases that are 'clean' and curated include UNITE for molecular identification of fungi (http://unite.ut.ee) and the Map of Life initiative (http://www.mappinglife.org) and their data providers. Data cleaning usually involves a combination of automated tools and manual expert curation. For instance, automated algorithms are now available to resolve common synonymisation issues, such as the Taxonomic Name

Resolution Service (TNRS; Boyle *et al*. 2013) and TaxoSaurus (Stoltzfus *et al*. 2013), but these rely on stable lists of accepted names for all species (e.g. http://www.catalogueoflife.org) – something that is fundamentally a taxonomic/nomenclatural challenge, not a technological issue. However, even if this would be achieved, it is now clear that taxonomic names cannot be used unambiguously, requiring biodiversity scientists to agree on and eventually fully adopt unique taxonomic identifiers (Kennedy *et al*. 2005).

THE PROMISES AND CHALLENGES OF BIODIVERSITY WORKFLOWS

The deluge of biological data and publications (e.g. Fig. 1) has been followed by a corresponding, albeit more modest, software development in ecology and evolution. This means that addressing relatively simple scientific questions may require researchers to master dozens of different analytical tools, often written in different programming languages and sometimes only available for select operating systems or programming environments. The complexity of the task increases as each tool is constantly updated, improved, and made more complex, or superseded by better methods. To tackle this problem, there is an increasing tendency to create integrative analytical platforms for ecological and evolutionary research. This is seen in a number of popular software packages, e.g. available in the R programming language (http://cran.r-project.org) and the Bio* toolkits in the Python, Ruby, Java, and Perl programming languages (http://open-bio.org), as well as stand-alone and on-line workflows (e.g. http://www.arborworkflows.com and http://www.biovel.eu).

There is no such thing as a 'standard analytical procedure' to investigate ecological and evolutionary processes. On the contrary, the choice of methodology will always depend on the research question, nature of the data, and the researcher's individual skills and knowledge to select and carry out analyses. This lack of procedural consensus might at first be perceived as problematic, but in reality it fuels scientific advancement and can be expected to occur in every step in a modern study of any noteworthy scope – from data acquisition to analysis and interpretation. In practice, it means that any bioinformatic platform to handle large amounts of biodiversity data needs to be highly modular and flexible. Researchers should be allowed to make their own choices concerning for instance the inclusion/exclusion of species, the choice of genetic markers, what fossils and methodology to employ for molecular dating, the delimitation of areas for biogeographic and diversification analyses, and what analytical tools to use.

The point of departure for any rigorous analysis should be that *all available data of adequate quality should be included* unless there are specific reasons to warrant the exclusion of parts of the data. It is, in fact, hard to justify why an ecological, phylogenetic, or biogeographic study of a given group (i.e. family or order) should not include all high-quality sequences, geo-referenced specimens, and fossil calibrations available for the group. Nevertheless, the great majority of analytical platforms currently available (e.g. the CIPRES gateway at http://www.phylo.org/portal2) require users to upload their own data for analysis, rather than providing datasets for the group or question of interest. The practical and theoretical challenges highlighted in this paper, and the time required for data assembly and cleaning, imply that

researchers are most likely missing valuable data, or even including data not properly assembled (e.g. without adequate assessments of gene orthology or verification of species distributions).

Modern biodiversity tools should thus tackle a moving target – the needs of the modern scientist interested in addressing crucial ecological and evolutionary questions in the face of rapid data growth and methodological development.

PRESENTING THE SUPERSMART APPROACH

Here we introduce a new conceptual and bioinformatic approach: SUPERSMART (*Self-Updating Platform for Estimating Rates of Speciation and Migration, Ages, and Relationships of Taxa*, http://www.supersmart-project.org). SUPERSMART aims at tackling the main problems outlined above by estimating continuously updated, time-calibrated molecular trees for all species, and providing researchers with a flexible, modular, integrative, and open-source platform for hypothesis-driven research in ecology and evolutionary biology. Thanks to a user-friendly platform, with only a few configuration steps any user will be able to either pull out custom-made sets of robustly inferred, dated trees for further analyses, or to assemble aligned DNA datasets representing any combination of sequenced genes/markers and taxa, and (in upcoming versions) along with data on their ecology, distribution, fossil records, and climatic characteristics. Our aim is that SUPERSMART will provide an *"integrative biogeography solution"* envisioned by Wen *et al*. (2013), ultimately aimed at elucidating the evolution of past

and present species distributions, guiding conservation efforts, and visualizing these data in a comprehensive way.

Overview. SUPERSMART consists of a set of programs that use the functionality provided by a common modular programming framework. The framework forms the bridge between the logical concepts that feature in the pipeline (sequences, taxa, fossils, and trees), the records in a relational database that contains local copies of a number of public resources, and the operations needed to assemble these records into tailored datasets and analyse them. The pipeline can be installed in environments that support the hosting and provisioning of free operating systems of the UNIX family. A web-based graphical user interface is planned, which will not require the local installation of any software.

In practice, a user infers arbitrary-sized, multi-marker, recursive phylogenies for large collections of species of interest (or one or more higher taxa) based on all suitable genes/genetic markers. An overview of how SUPERSMART relates to available supermatrix and supertree approaches is shown in Fig. 2. In SUPERSMART, the included genetic markers may typically comprise DNA barcodes (Hebert *et al*. 2003), i.e. COI, *rbcL*, *matK,* and ITS, as well as select markers known to perform particularly well in specific groups of organisms, including those generated with next-generation sequencing techniques. To enable the inclusion of arbitrarily many species in the final results – potentially spanning the whole tree of life – we employ a *three-step approach* (Fig. 3):

1. We initially build a backbone, higher-level tree that features an optimally distributed set of broadly sequenced exemplar species. This backbone tree (a phylogram) is time-calibrated using all suitable fossils from a fossil calibration table (see below for details);
2. We decompose this 'mega-chronogram' into subclades (typically equivalent to families or genera) that are strongly supported and contain a manageable number of descendent species. All descendant taxa and their genetic markers are added to these subclades. By default, SUPERMART selects one terminal per species, but users may chose to include all intra-specific taxa down to the level of individuals. Time-calibrated species trees are then inferred under the multi-species, multi-marker coalescent model (as implemented in *BEAST). These trees are re-calibrated (scaled) to the posterior age obtained for the clade in Step 1;
3. The dated species-level trees are then grafted into a complete species-level chronogram of directly comparable ages. These trees can now be used for various post-tree spatial and temporal analyses, including inferrences of e.g. migration, diversification, and niche evolution (some of which are already implemented at http://www.biovel.eu and are planned to be integrated with SUPERSMART).

Data mining. SUPERSMART mines public databases for suitable DNA sequences by way of their globally unique taxonomic identifiers. Our present proof of concept adopts those defined by Federhen (2012), but this could be extended to recognize other unambiguous identifiers, such as internet addresses or uniform resource identifiers (URIs). All species names the user is interested

in are mapped onto such unambiguous identifiers, before downstream analyses take place. This is accomplished in one of two ways: *i)* by the user preparing and providing a mapping that contains, for each name of interest, the corresponding identifier; or *ii)* by an automated step that prepares such a mapping, which is done by querying the *Taxosaurus* service (http://taxosaurus.org) for each user-supplied input name and attempting to resolve it, accounting for synonyms and misspellings in the process.

SUPERSMART compiles plausible candidate sets of DNA sequences for alignment, orthology assessment, and subsequent phylogenetic inference by querying a local, modified version of the PhyLoTA database (Sanderson 2008; Sanderson *et al*. 2008). In brief, this database is the product of a process that crawls all taxonomically organized GenBank sequence divisions and at certain strategically chosen nodes performs all-versus-all similarity searches of the sequences subtended by that node. The sets of search hits are then grouped into single linkage clusters. Due to the care with which the PhyLoTA browser has optimized the search parameters used in this process, the sequences that are grouped in these clusters are generally a good starting point for phylogenetic inference, although several further data processing steps are necessary. These are described below.

Data reduction. Many PhyLoTA clusters contain multiple sequences from the same species, often with extensive sampling bias towards "model organisms" (*sensu* PhyLoTA). As the standard goal of SUPERSMART is to infer species-level trees (although lower taxonomic levels are also supported), these masses of sequences must be reduced (dereplicated) to more

manageable datasets, containing approximately equal numbers of sequences for each species. The current approach is therefore to select the most complete sequences, i.e. the one with the fewest DNA ambiguity symbols (Cornish-Bowden 1985) and that most closely approaches the median length of all sequences for that species in that cluster. The goal is to avoid overly short sequence fragments for markers for which longer stretches are readily available for the same species. In addition, we also want to avoid considerably longer stretches where the marker has been sequenced as part of, e.g. a mitochondrial or chloroplast genome. Even though instances of either scenario are generally avoided due to the minimum requirements that PhyLoTA imposes on overlap of reciprocal hits, these additional steps make SUPERSMART produce even cleaner datasets – which are more representative for intra-specific sequence variation and contain less missing data.

Data merging. PhyLoTA clusters consist of sets of putatively homologous sequences grouped at a taxonomic level deemed appropriate (depending on size these may be e.g. genera, families, or orders). Therefore, multiple "sister clusters" may exist for the same marker. For SUPERSMART to infer phylogenies that span several of these low taxonomic levels, such sister clusters need to be merged correctly. There is no consensus in the scientific community on how to tackle this issue, but we have identified three possible approaches:

1. to make use of curated annotations of seed sequence metadata such that all candidate clusters whose seed sequence is annotated are merged into a single alignment, e.g. for *rbc*L or COI. This would be a feasible option for a limited number of markers with standardized naming conventions (such as DNA barcodes). However, our evaluation

of the heterogeneity of names applied to non-standardized markers suggests that this is an impractical approach for most other markers, despite community initiatives to address this (e.g. see http://genenames.org);

2. to identify orthology among candidate sister clusters by searching for the protein translation of the seed sequence of each candidate cluster against the InParanoid database (O'Brien *et al*. 2005). InParanoid assigns orthology for all protein-coding genes in the genomes of a large number of model organisms. If the best hits of queried sequences are each other's orthologs according to InParanoid, then, transitively, so are the PhyLoTA clusters to which the query sequences belong. This approach, however, can only be applied to protein-coding genes;
3. to run all-vs-all similarity searches among the set of seed sequences that represent the clusters that are candidates for merging. In essence, this is the procedure also used by PhyLoTA to select higher taxonomic levels to form "super clusters". This has proved to be the most successful approach, and one that can be applied also to non-coding regions as well as regions that lack standardized names. It is therefore included in the standard implementation of SUPERSMART.

Multiple sequence alignment. The DNA sequences as stored in our database are unaligned. As merged clusters can ultimately grow to very large numbers of sequences, we designed the pipeline in such a way that multiple sequence alignment takes place as a two-step process. First, the clusters as assigned by PhyLoTA are aligned (after data reduction). Second, orthology among clusters at taxonomic levels higher than PhyLoTA can manage is identified using the approach

492 described above and "sister clusters" are subsequently merged using profile alignment, as is also
493 the case with the Phylogenetic Dataset Construction toolkit (PHLAWD; http://phlawd.net/). By
494 default, the first alignment step uses MAFFT (Katoh & Standley 2013) and the second, profile
495 alignment, uses MUSCLE (Edgar 2004), although the system can be configured to use any of a
496 range of other alignment programs. Wrappers are provided for several other alignment programs.
497

498 ***Marker and taxon selection.*** The data compilation steps outlined above provide a wealth of data,
499 although not all of them may be suitable for producing a backbone tree. An optimal balance must
500 be found between taxon sampling, taxon overlap, sequence divergence, and overall size (and
501 sparseness) of the combined data. In our multi-step approach, this optimum is further influenced
502 by which exemplar species are selected for the backbone inference step. Researchers who have
503 not attempted to assemble similar datasets may probably not realise the non-triviality of the task,
504 which is directly related to the classical *"knapsack problem"* (Martello & Toth 1990) (Fig. 4A).
505 Our approach for exemplar selection is to select, for each higher taxon to be represented,
506 the two species that most frequently form the most distal pair when computing all pairwise
507 sequence distances within the higher taxon. This is done iteratively for each candidate alignment.
508 During this step we weigh the occurrence of distal pairs by n-1, where n is the number of
509 pairwise comparisons within each alignment. The rationale is that the most distal pair among a
510 large number of comparisons is more likely to "cross the root" of the containing higher taxon (or
511 at least, represent a deep split) than in smaller alignments. For the trivial case of an alignment
512 only including two species, the pairwise distance between them is consequently (n-1=0) and thus
513 discarded as uninformative.

Once all exemplar species are identified, candidate alignments are selected for concatenation as input in the backbone analysis. For this step the user can define a maximum amount of average pairwise sequence divergence (to prevent the inclusion of saturated alignments) and a minimum number of alignments within which each species must occur. We then attempt to solve the *"knapsack problem"* of packing the required number of suitable alignments into a minimally sparse supermatrix. The greedy approximation approach we take (Fig. 4B) is to sort the exemplar species in increasing order of participation in candidate alignments (i.e. rarely sequenced species are treated first). We then sort the alignments in decreasing taxon coverage. Finally, we iteratively visit the species and for each of them, add its available alignments to the supermatrix, until the focal species' minimum participation threshold has been breached. During this process we increment the occurrence counts for all other species that also occur in the alignments that are added to the supermatrix, so that frequently sequenced species have likely already exceeded their minimum occurrence threshold without requiring separate treatment.

Using the supermatrix of concatenated alignments for the exemplar species, we then infer a backbone phylogeny. Given that the supermatrix may span several thousand taxa, we employ highly scalable tree inference methods, providing end users with a choice between ExaML (Stamatakis & Aberer 2013), which is based on a maximum likelihood algorithm, and ExaBayes (http://sco.h-its.org/exelixis/web/software/exabayes) based on Bayesian inference.

Time calibration using fossils. Until the Fossil Calibration Database linked to Paleontologica Electronica (Ksepka *et al*. 2011; Parham *et al*. 2012b) becomes fully operational and taxonomically well-sampled, we have created a provisional database of fossil calibration points. In addition, we introduce an index termed *'best practice score'* calculated based on the five criteria set up by Parham et al. (2012a). One advantage of this reliability index is to allow the user to decide whether to use only the most expertly assessed (but fewer) fossils as calibrations, or to experiment with additional fossils of lower confidence. These may include fossils whose phylogenetic placement may be more tentative, but which might better inform the true age of the calibrated nodes (Sauquet *et al*. 2012). This database is a cloned subset of a module implemented in PROTEUS (http://eflower.myspecies.info/proteus), which offers an ideal setting for entering data, keeping track of changes (which, by whom, and based on what), and data-mining. An example of a fossil calibration record in this module is shown in Supplementary Fig. S1.

On the resulting backbone tree inferred in the previous step, SUPERSMART then maps all suitable fossils belonging to the focal clade, as directly exported from PROTEUS. This is the stage at which fossils can be filtered based on a user-defined reliability score. The tree is then dated using the relaxed clock algorithm Penalized Likelihood (Sanderson 2002) further developed and implemented in treePL (Smith & O'Meara 2012), which can handle very large numbers of terminals. Other popular dating tools such as BEAST (Drummond & Rambaut 2007) are not yet feasible for more than a few hundred taxa. In upcoming versions, this module is planned to handle samples of trees as input, and thereby produce confidence intervals of node ages rather than point estimates.

Species-level analyses. Using the backbone topology, SUPERSMART assesses whether the selected exemplar species are recovered as monophyletic pairs. If this is not the case, it traverses up the backbone until a larger assemblage is formed that is, in total, monophyletic with respect to all outgroup taxa (although its members may be paraphyletic with respect to each other). For these assemblages we then select all available alignments. The user is provided the option to define a maximum for the average pairwise sequence divergence within each alignment (to avoid saturated alignments) and a minimum amount of alignment density, i.e. the minimum fraction of the total number of species in the assemblage that must be represented in the candidate alignment to warrant inclusion. The set of alignments selected for the focal assemblage of species is then analysed under the multi-species, multi-marker coalescent implemented in *BEAST (Heled & Drummond 2010).

The resulting ultrametric species-level subtree is then grafted back onto the backbone chronogram. First, all branch lengths on the subtree are re-scaled such that the most recent common ancestor of the exemplars in the subtree is set to the same age-before-present (distance to the tips) as the equivalent node in the backbone tree. As both trees are ultrametric, this distance can be directly compared. If the exemplar species in the subtree are on either side of the root, then the pair of exemplars in the backbone can simply be replaced by the subtree. If not, then the distance between the most recent common ancestor of the exemplars in the subtree and the root of the subtree is computed. This difference is then subtracted from the branch leading up to the exemplar pair in the backbone, and from that point onwards the subtree is grafted in place

of the exemplar pair. The result is a species-level dated phylogeny with directly comparable clade ages, including all suitable species and genetic markers publicly available, and any additional data provided by the user – all done with no more than a few clicks.

EMPIRICAL EXAMPLES

We present the functionality of SUPERSMART in its current implementation on two empirical datasets: the mammalian order Primates and the plant order Gentianales. These taxa provide contrasting examples commonly encountered in eco-evolutionary research. Primates have been extensively studied by the scientific community, leading to a massive accumulation of sequences, which are however highly biased towards our own species and near relatives. Even so, the estimated number of living species range from 249 (http://www.catalogueoflife.org) to 376-450 (Springer *et al*. 2012), showing how the classification of even such a charismatic clade remains a topic of debate. The Gentianales, despite comprising a much larger number of extant species (c. 22,237 according to Catalogue of Life) and several economically important genera such as *Coffea*, *Catharanthus, Cinchona,* and *Strychnos*, have received considerably less attention and are therefore the subject of much lower genetic coverage in public sequence databases. Table 1 shows summary statistics for these two taxa as well as core parameter values used during Steps 1 and 2 of their SUPERSMART analyses.

Gentianales. Figure 5 shows the final time-calibrated species level tree comprising 701 species in all five families of the order: Apocynaceae, Gentianaceae, Rubiaceae, Loganiaceae and

Gelsemiaceae (see Fig. S2 for a fully annotated tree). The Step 2 analysis included as many as 54 different genetic markers, thus providing the hitherto most comprehensive species-level analysis of the order to date. The group is estimated to have originated c. 103 million of years ago (Ma). All families form distinct monophyletic groups, and their relationship corroborates previous phylogenetic estimates, e.g. showing Rubiaceae as sister to the rest of the order (e.g. Backlund *et al*. 2000; Jiao & Li 2007). Higher-level relationships within the Rubiaceae, the largest family with 13,514 species, are also in large agreement with the recent family-level phylogenetic analysis of Bremer & Eriksson (2009): subfamilies Cinchonoideae, Ixoroideae, and Rubioideae are found to be monophyletic, with Rubioideae as sister to the others. Furthermore, inter-tribal relationships within Rubioideae are largely consistent with their findings.

Primates. Figure 6 shows the results from the analysis of this order using another way of visualising large trees, while figure S3 presents the fully annotated species tree comprising 178 species. The topology of the extant lineages Strepsirrhini (78 Ma), Tarsiers (69 Ma), New World monkeys (50 Ma), Old World monkeys, and apes (33 Ma) is correctly reconstructed. The phylogeny is based on more than 20 markers (see Table 1), of which three had to be available for a species to be included in Step 1 of the analyses. The relationships among the families within the New world monkeys (*Platyrrhini*) are still unclear (Opazo *et al*. 2006), but all genera are supported here as monophyletic. Our results suggest an initial split of the family *Pitheciidae* and a close relationship between the families *Atelidae* and *Cebidae*. The Old World monkeys (*Cercopithecidae*) comprise the two monophyletic subfamilies *Colobinae* and *Cercopithecinae*

which is in agreement with previously published primate trees (Vos 2006). The Hominoidea is well resolved, including the resolution of the hominoid trichotomy. Relatively less data were available for the inference of the *Taarsiformes* and the *Strepsirrhini*. The *Strepsirrhini* split into Malagasy and non-Malagasy species. The lemurs, native to Madagascar, are represented by four families that are well resolved in our tree. However, due to low sequencing data coverage, many *Strepsirrhini* genera are represented by only their two respective exemplar species.

Interactions with other initiatives. Figure 7 outlines some of the anticipated interactions and data exchange during different operational levels. SUPERSMART is designed as a community-based platform, which will complement and interact (rather than compete) with many ongoing initiatives worldwide. For instance, a related application is the Phylogenetic Dataset Construction toolkit (PHLAWD; http://phlawd.net/), which efficiently assembles sequence data for a pre-specified list of target species (Zanne *et al*. 2014). SUPERSMART has similar goals but differs from PHLAWD by dealing more extensively with name resolution, homology assessment, optimal taxonomic vs. genetic coverage, and time-calibration through a native curated fossil table and plugged-in tools, among other differences in scope and functionality. Two other recent projects have similarities with SUPERSMART, although their scope is more limited. PUmPER (Izquierdo-Carrasco *et al*. 2014) assembles multiple sequence alignments for a given group in the NCBI taxonomy, but it constructs maximum likelihood gene trees, omitting the Bayesian multi-species, multi-marker coalescent approach we present, the time calibration step, and the outlook towards integration with other data by way of taxonomic name resolution. In addition, the PUmPER approach is designed to construct a single tree, which may pose

scalability problems for very large numbers of taxa, in contrast with the recursive approach we present here. Huerta-Cepas *et al*. (2014) present such a recursive approach through their Nested Phylogenetic Reconstruction (NPR) methodology, but the end result, like PUmPER, is a maximum likelihood gene tree whose branch lengths are not proportional to time. Crucially, in both cases, the resulting estimate of phylogeny is not an ultrametric species tree, which notably hampers their application in phylogenetic, diversification, and ecological comparative analyses.

On-going developments. The version of SUPERSMART released with this publication contains a fully functional set of tools for performing Steps 1 and 2 described above (Fig. 3). In the next version, we will implement several post-tree tools (i.e., Step 3). These will include mapping distribution data into phylogenies (SpeciesGeoCoder; https://github.com/mtop/geocoder), extracting subtrees from global species-level chronograms (http://phylotastic.org), estimating area-specific migration and diversification rates under a Bayesian framework (FitzJohn 2010; Silvestro *et al*. 2011), and integrating the SUPERSMART tools with the workflows at BioVeL (http://www.biovel.eu). Many enhancements are planned and will be successively added, including visually appealing and flexible Graphical User Interfaces (GUI). Anyone can join the users' list and request additional features, and those wishing to contribute to the code and project may also request to join the developers' list at https://github.com/naturalis/supersmart.

CONCLUSION AND PROSPECTS

Biological research has arguably never been as exciting – but also as challenging – as today. We have entered the era of Big Data and cannot ignore its potential impact on the questions we address. As an evolutionary perspective (e.g. phylogenetic signal) is increasingly acknowledged as an essential component of ecological studies at various scales (Srivastava *et al*. 2012; Chave 2013), integrative bioinformatic solutions such as SUPERSMART will aid researchers to tackle the 'moving target' of data accumulation, methodological development, and theoretical advances.

AVAILABILITY

All source code underlying this project is freely accessible under an MIT license at http://www.supersmart-project.org and http://antonelli-lab.net, where tutorials, example files, and other relevant information will be continuously made available. Access to the fossil calibration module in PROTEUS may be requested at http://eflower.myspecies.info/proteus.

ACKNOWLEDGEMENTS

We thank Johan Nylander, Matthias Obst, Christine Bacon, Elisabet Sjökvist, Cajsa Lisa-Anderson, Carlos Jaramillo, Klas Benjaminsson, and various other colleagues for discussions and support; Luke Harmon, Karen Cranston, Daniel Ksepka, Walter Jetz, James Rosindell, for discussions on the integration of our initiatives. Funding was provided by the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013, ERC Grant Agreement n. 331024) to A.A.; the Swedish Research Council to A.A. and B.O.; project BioVeL grant no. 283359 to H.H.; Carl Tryggers stiftelse (CTS 12:24, 11:479 and 12:507) to

F.L.C., D.S. and M.T.; Wenner-Gren to D.S.; FORMAS (215-2011-498) to R.H.N. The code was developed, tested, and benchmarked on the bioinformatics computer cluster Albiorix at the Department of Biological and Environmental Sciences, University of Gothenburg, and the HPC infrastructure at Naturalis Biodiversity Center.

FIGURES

Fig 1. Biodiversity research has entered the era of Big Data. **A:** Growth of the number of scientific publications in ecology and evolution (search for subject headings in the PubMed database). **B:** Snapshot of species occurrence records in the Global Biodiversity Information Facility. **C:** Fossil occurrences available from the Paleobiology Database. **D:** Growth of molecular data deposited in GenBank.

Fig 2. Methods for inferring large (dated) phylogenies. Schematic comparison of the supertree, supermatrix, and the SUPERSMART approaches (presented in this paper).

Fig. 3. Basic overview of SUPERSMART. **Step 1**: a hypothetical, dated backbone phylogenetic tree comprising 16 genera, each represented by a single species. The dating employs all available fossils suitable for time calibration. **Step 2** depicts a clade (2 genera with 11 species in total) for which no direct fossil calibration is available. Divergence times can nevertheless be confidently estimated by re-calibrating (scaling) the branches with the posterior clade ages obtained in Step 1. **Step 3:** The fine-level, dated trees are then grafted together and made available for various post-tree analyses in ecology and evolution, relying on directly comparable divergence times.

Fig. 4. Illustration of the classic knapsack problem, applied to the optimal choice of species and alignments (markers) for compiling DNA alignments. **A:** Combinatorial problem of maximizing the amount of money in the knapsack with a maximum capacity of 15 kg. **B:** Seven exemplar species (S1-S7) are put in ascending order by their occurrence in the candidate alignments (A1-

A7) which are in turn ordered by taxon coverage. In this example, the minimum number of alignments per species is set to two. The supermatrix is then compiled as described in the text. The resulting matrix consists of five alignments and only six species, since the number of alignments in which species S4 occurs does not meet the required minimum.

Fig. 5. Time-calibrated phylogeny of the plant order Gentianales constructed using SUPERSMART, including 701 species and calibrated using 10 fossils. The five families are outlined. Internal concentric circles represent 10 Ma bins. See Fig. S2 for a fully annotated tree. The figure was generated with FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Fig. 6. Molecular chronogram of Primates constructed with SUPERSMART, shown in the tree visualisation tool OneZoom (Rosindell & Harmon 2012; http://www.onezoom.org/). Top left: Full tree; top right: zoom to the genus *Saguinus*; bottom: zoom to a branch comprising four species. An interactive image is available at http://www.supersmart-project.org/p/example.html. See also Fig. S3 for the fully annotated tree in classical style.

Fig. 7. Proposed interactions with other initiatives at different analytical stages. *"Global analyses"* will provide continuously updating, dated phylogenies of all species with publicly available molecular sequences, and (in upcoming versions) estimates of diversification and migration rates among and within a set of pre-defined GIS polygons (such as WWF's realms and biomes). The results may be retrieved by other initiatives and will be deposited in data repositories. *"User-defined analyses"* are influenced by individual choices, including user-

defined polygons (areas), taxa of interest, and fossil records. The user may also include data that are not yet published or are not public. A series of post-tree analyses will be available, mainly through BioVeL workflows, and new ones will be incorporated continuously.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.501v1 | CC-BY 4.0 Open Access | rec: 15 Sep 2014, publ: 15 Sep 2014

ONLINE SUPPLEMENTARY MATERIAL

Supplementary Fig. S1. Example of a fossil entry in the PROTEUS database. Relevant information such as the taxon name (here *Endressinia brasiliana*), age boundaries, and 'best practice scores' (see text for explanation) can be exported and directly used for time calibration in SUPERSMART.

Supplementary Fig. S2. Phylogeny of Gentianales. The final tree inferred using the SUPERSMART pipeline comprises the five families *Apocynaceae* (blue), *Gentinaceae* (yellow), *Rubiaceae* (red), *Loganiaceae* (green), and *Gelsemiaceae* (turquoise). Node labels and branch widths represent posterior support values for inferences of individual clades. The figure was generated with FigTree.

Supplementary Fig. S3. Fully annotated phylogeny of Primates inferred using SUPERSMART. Node labels represent posterior support values for inferences of individual clades. The figure was generated with FigTree.

TABLES

Table 1. Summary statistics for SUPERSMART Steps 1 and 2 of the orders Gentianales and Primates.

764 **REFERENCES**

765 1.

766 Aberer, A.J., Krompass, D. & Stamatakis, A. (2013). Pruning Rogue Taxa Improves
767 Phylogenetic Accuracy: An Efficient Algorithm and Webservice. *Syst. Biol.*, 62, 162-166.
768 2.

769 Bayzid, M.S. & Warnow, T. (2012). Estimating Optimal Species Trees from Incomplete Gene
770 Trees Under Deep Coalescence. *J. Comput. Biol.*, 19, 591-605.
771 3.

772 Bengtsson-Palme, J., Ryberg, M., Hartmann, M., Branco, S., Wang, Z., Godhe, A. *et al.* (2013).
773 Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS
774 sequences of fungi and other eukaryotes for analysis of environmental sequencing data.
775 *Methods in Ecology and Evolution*, 4, 914-919.
776 4.

777 Bidartondo, M., Bruns, T.D., Blackwell, M., Edwards, I., Taylor, A.F., Horton, T. *et al.* (2008).
778 Preserving accuracy in GenBank. *Science*, 319.
779 5.

780 Boakes, E.H., McGowan, P.J., Fuller, R.A., Chang-qing, D., Clark, N.E., O'Connor, K. *et al.*
781 (2010). Distorted views of biodiversity: spatial and temporal bias in species occurrence
782 data. *PLoS Biol.*, 8, e1000385.
783 6.

784 Boyle, B., Hopkins, N., Lu, Z., Raygoza Garay, J.A., Mozzherin, D., Rees, T. *et al.* (2013). The
785 taxonomic name resolution service: an online tool for automated standardization of plant
786 names. *BMC Bioinformatics*, 14, 16.
787 7.

788 Chave, J. (2013). The problem of pattern and scale in ecology: what have we learned in
789 20 years? *Ecol. Lett.*, 16, 4-16.
790 8.

791 Chesters, D. & Vogler, A.P. (2013). Resolving Ambiguity of Species Limits and Concatenation
792 in Multilocus Sequence Data for the Construction of Phylogenetic Supermatrices. *Syst.*
793 *Biol.*, 62, 456-466.
794 9.

795 Cody, S., Richardson, J.E., Rull, V., Ellis, C. & Pennington, R.T. (2010). The great American
796 biotic interchange revisited. *Ecography*, 33, 326-332.
797 10.

Cornish-Bowden, A. (1985). Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res*., 13, 3021.
11.

Crisp, M.D., Arroyo, M.T.K., Cook, L.G., Gandolfo, M.A., Jordan, G.J., McGlone, M.S. *et al*. (2009). Phylogenetic biome conservatism on a global scale. *Nature*, 458, 754-756.
12.

de Mendonça, R.S., Navia, D., Diniz, I.R., Auger, P. & Navajas, M. (2011). A critical review on some closely related species of Tetranychus sensu stricto (Acari: Tetranychidae) in the public DNA sequences databases. *Exp. Appl. Acarol*., 55, 1-23.
13.

de Queiroz, A. & Gatesy, J. (2007). The supermatrix approach to systematics. *Trends in Ecology and Evolution*, 22, 34-41.
14.

Drew, B.T. (2013). Data deposition: Missing data mean holes in tree of life. *Nature*, 493, 305-305.
15.

Drew, B.T., Gazis, R., Cabezas, P., Swithers, K.S., Deng, J., Rodriguez, R. *et al*. (2013). Lost Branches on the Tree of Life. *PLoS Biol*, 11, e1001636.
16.

Drummond, A.J., Ho, S.Y., Phillips, M.J. & Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol*., 4.
17.

Drummond, A.J. & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol*., 7, 8.
18.

Duputié, A., Zimmermann, N.E. & Chuine, I. (2014). Where are the wild things? Why we need better data on species distribution. *Global Ecol. Biogeogr*., 23, 457-467.
19.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*., 32, 1792-1797.
20.

Edwards, S.V. (2009). Is a new and general theory of molecular systematics emerging? *Evolution*, 63, 1-19.
21.

Federhen, S. (2012). The NCBI taxonomy database. *Nucleic Acids Res*, 40, D136-D143.

22.

FitzJohn, R.G. (2010). Quantitative Traits and Diversification. *Syst. Biol.*, 59, 619-633.
23.

Fujita, M.K., Leaché, A.D., Burbrink, F.T., McGuire, J.A. & Moritz, C. (2012). Coalescent-based species delimitation in an integrative taxonomy. *Trends in ecology & evolution (Personal edition)*, 27, 480-488.
24.

Gernhard, T., Ford, D., Vos, R. & Steel, M. (2006). Estimating the relative order of speciation or coalescence events on a given phylogeny. *Evolutionary bioinformatics online*, 2, 285.
25.

Goloboff, P.A., Catalano, S.A., Marcos Mirande, J., Szumik, C.A., Salvador Arias, J., Källersjö, M. *et al*. (2009). Phylogenetic analysis of 73 060 taxa corroborates major eukaryotic groups. *Cladistics*, 25, 211-230.
26.

Gotelli, N.J. & Colwell, R.K. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology letters*, 4, 379-391.
27.

Guralnick, R.P., Wieczorek, J., Beaman, R., Hijmans, R.J. & the BioGeomancer Working, G. (2006). BioGeomancer: Automated Georeferencing to Map the World's Biodiversity Data. *PLoS Biol*, 4, e381.
28.

Gustafsson, A.L.S., Verola, C.F. & Antonelli, A. (2010). Reassessing the temporal evolution of orchids with new fossils and a Bayesian relaxed clock, with implications for the diversification of the rare South American genus Hoffmannseggella (Orchidaceae: Epidendroideae). *BMC Evol. Biol.*, 10, 177.
29.

Hebert, P.D.N., Cywinska, A., Ball, S.L. & DeWaard, J.R. (2003). Biological identifications through DNA barcodes. *Proc. R. Soc. Lond., Ser. B: Biol. Sci.*, 270, 313-321.
30.

Heled, J. & Drummond, A.J. (2010). Bayesian Inference of Species Trees from Multilocus Data. *Mol. Biol. Evol.*, 27, 570-580.
31.

Ho, S.Y.W. & Phillips, M.J. (2009). Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Syst. Biol.*, 58, 367-380.
32.

Hoorn, C., Wesselingh, F.P., ter Steege, H., Bermudez, M.A., Mora, A., Sevink, J. *et al*. (2010). Amazonia Through Time: Andean Uplift, Climate Change, Landscape Evolution, and Biodiversity. *Science*, 330, 927-931.
33.

Huerta-Cepas, J., Marcet-Houben, M. & Gabaldón, T. (2014). A nested phylogenetic reconstruction approach provides scalable resolution in the eukaryotic Tree Of Life. PeerJ PrePrints.
34.

Hyde, K.D., Udayanga, D., Manamgoda, D.S., Tedersoo, L., Larsson, E., Abarenkov, K. *et al*. (2013). Incorporating molecular data in fungal systematics: a guide for aspiring researchers. *arXiv preprint arXiv:1302.3244*.
35.

Izquierdo-Carrasco, F., Cazes, J., Smith, S.A. & Stamatakis, A. (2014). PUmPER: phylogenies updated perpetually. *Bioinformatics*, 30, 1476-1477.
36.

Johnsrud, S., Yang, H.G., Nayak, A. & Punyasena, S.W. (2013). Semi-automated segmentation of pollen grains in microscopic images: a tool for three imaging modes. *Grana*, 52, 181-191.
37.

Jones, K.E., Bielby, J., Cardillo, M., Fritz, S.A., O'Dell, J., Orme, C.D.L. *et al*. (2009). PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals: Ecological Archives E090-184. *Ecology*, 90, 2648-2648.
38.

Katoh, K. & Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, 30, 772-780.
39.

Kennedy, J.B., Kukla, R. & Paterson, T. (2005). Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration. In: *Data integration in the life sciences*. Springer, pp. 80-95.
40.

Ksepka, D.T., Benton, M.J., Carrano, M.T., Gandolfo, M.A., Head, J.J., Hermsen, E.J. *et al*. (2011). Synthesizing and databasing fossil calibrations: divergence dating and beyond. *Biol. Lett.*, 7, 801-803.
41.

Kupczok, A., Schmidt, H. & von Haeseler, A. (2010). Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. *Algorithms Mol Biol*, 5, 37.

42.

Linder, H.P., Hardy, C.R. & Rutschmann, F. (2005). Taxon sampling effects in molecular clock dating: An example from the African Restionaceae. *Mol. Phylogen. Evol.*, 35, 569-582.

43.

Martello, S. & Toth, P. (1990). *Knapsack problems*. Wiley New York.

44.

Nakamura, Y., Cochrane, G., Karsch-Mizrachi, I. & Database, I.N.S. (2013). The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.*, 41, D21-D24.

45.

Nguyen, N., Mirarab, S. & Warnow, T. (2012). MRL and SuperFine+MRL: new supertree methods. *Algorithms for Molecular Biology*, 7, 3.

46.

O'Brien, K.P., Remm, M. & Sonnhammer, E.L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, 33, D476-D480.

47.

Opazo, J.C., Wildman, D.E., Prychitko, T., Johnson, R.M. & Goodman, M. (2006). Phylogenetic relationships and divergence times among New World monkeys (Platyrrhini, Primates). *Mol Phylogenet Evol*, 40, 274-280.

48.

Pamilo, P. & Nei, M. (1988). Relationships between gene trees and species trees. *Mol Biol Evol*, 5, 568-583.

49.

Parham, J.F., Donoghue, P.C.J., Bell, C.J., Calway, T.D., Head, J.J., Holroyd, P.A. *et al*. (2012a). Best Practices for Justifying Fossil Calibrations. *Syst. Biol.*, 61, 346-359.

50.

Parham, J.F., Ksepka, D.T., Polly, P.D., Van Tuinen, M. & Benton, M.J. (2012b). The Fossil Calibration Database: A New Bioinformatic Tool for Dating Divergences of Extant Lineages by Synthesizing Paleontological and Molecular Sequence Data. *J. Vert. Paleontol.*, 32, 154-154.

51.

Price, M.N., Dehal, P.S. & Arkin, A.P. (2010). FastTree 2-Approximately Maximum-Likelihood Trees for Large Alignments. *Plos One*, 5.

52.

Rabosky, D.L., Slater, G.J. & Alfaro, M.E. (2012). Clade Age and Species Richness Are Decoupled Across the Eukaryotic Tree of Life. *PLoS Biol*, 10, e1001381.

53.

Rannala, B. & Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164, 1645-1656.
54.

Rosindell, J. & Harmon, L. (2012). OneZoom: A fractal explorer for the tree of life. *PLoS Biol.*, 10, e1001406.
55.

Sanderson, M.J. (2002). Estimating Absolute Rates of Molecular Evolution and Divergence Times: A Penalized Likelihood Approach. *Mol. Biol. Evol.*, 19, 101-109.
56.

Sanderson, M.J. (2003). r8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, 19, 301-302.
57.

Sanderson, M.J. (2008). Phylogenetic signal in the eukaryotic tree of life. *Science*, 321, 121-123.
58.

Sanderson, M.J., Boss, D., Chen, D., Cranston, K.A. & Wehe, A. (2008). The PhyLoTA Browser: Processing GenBank for molecular phylogenetics research. *Syst. Biol.*, 57, 335-346.
59.

Sauquet, H., Ho, S.Y.W., Gandolfo, M.A., Jordan, G.J., Wilf, P., Cantrill, D.J. *et al.* (2012). Testing the impact of calibration on molecular divergence times using a fossil-rich group: the case of Nothofagus (Fagales). *Syst. Biol.*, 61, 289-313.
60.

Silvestro, D., Schnitzler, J. & Zizka, G. (2011). A Bayesian framework to estimate diversification rates and their variation through time and space. *BMC Evol. Biol.*, 11, 311.
61.

Smith, S., Beaulieu, J. & Donoghue, M. (2009). Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. *BMC Evol. Biol.*, 9, 37.
62.

Smith, S.A., Brown, J.W. & Hinchliff, C.E. (2013). Analyzing and Synthesizing Phylogenies Using Tree Alignment Graphs. *PLoS Comput Biol*, 9, e1003223.
63.

Smith, S.A. & O'Meara, B.C. (2012). treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics*, 28, 2689-2690.
64.

Springer, M.S., Meredith, R.W., Gatesy, J., Emerling, C.A., Park, J., Rabosky, D.L. *et al*. (2012). Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. *Plos One*, 7, e49521.
65.

Srivastava, D.S., Cadotte, M.W., MacDonald, A.A.M., Marushia, R.G. & Mirotchnick, N. (2012). Phylogenetic diversity and the functioning of ecosystems. *Ecology letters*, 15, 637-648.
66.

Stamatakis, A. & Aberer, A.J. (2013). Novel Parallelization Schemes for Large-Scale Likelihood-based Phylogenetic Inference. In: *Parallel & Distributed Processing (IPDPS), 2013 IEEE 27th International Symposium on*, pp. 1195-1204.
67.

Stamatakis, A., Aberer, A.J., Goll, C., Smith, S.A., Berger, S.A. & Izquierdo-Carrasco, F. (2012). RAxML-Light: a tool for computing terabyte phylogenies. *Bioinformatics*, 28, 2064-2066.
68.

Stamatakis, A., Göker, M. & Grimm, G.W. (2010). Maximum likelihood analyses of 3,490 rbcl sequences: Scalability of comprehensive inference versus group-specific taxon sampling. *Evolutionary Bioinformatics*, 2010, 73-90.
69.

Stoltzfus, A., Lapp, H., Matasci, N., Deus, H., Sidlauskas, B., Zmasek, C.M. *et al*. (2013). Phylotastic! Making tree-of-life knowledge accessible, reusable and convenient. *BMC Bioinformatics*, 14, 158.
70.

Valkiunas, G., Atkinson, C.T., Bensch, S., Sehga, R.N.M. & Ricklefs, R.E. (2008). Parasite misidentifications in GenBank: how to minimize their number? *Trends Parasitol*., 24, 247-248.
71.

von Haeseler, A. (2012). Do we still need supertrees? *BMC Biol*., 10, 13.
72.

Vos, R.A. (2006). Inferring large phylogenies: the big tree problem. Ph.D. Thesis, Dept. of Biological Sciences, Simon Fraser University, Burnaby, B.C., Canada.
73.

Vos, R.A. & Mooers, A.Ø. (2004). Reconstructing divergence times for supertrees. In: *Phylogenetic Supertrees*. Springer Netherlands, pp. 281-299.
74.

Wen, J., Ree, R.H., Ickert-Bond, S.M., Nie, Z. & Funk, V. (2013). Biogeography: Where do we go from here? *Taxon*, 62, 912-927.
75.

Werdelin, L. & Sanders, W.J. (2010). *Cenozoic mammals of Africa*. University of California Press.
76.

Whidden, C., Zeh, N. & Beiko, R.G. (2014). Supertrees Based on the Subtree Prune-and-Regraft Distance. *Syst. Biol.*, 63, 566-581.
77.

Wiens, J.J. (1998). Does adding characters with missing data increase or decrease phylogenetic accuracy? *Syst. Biol.*, 47, 625-640.
78.

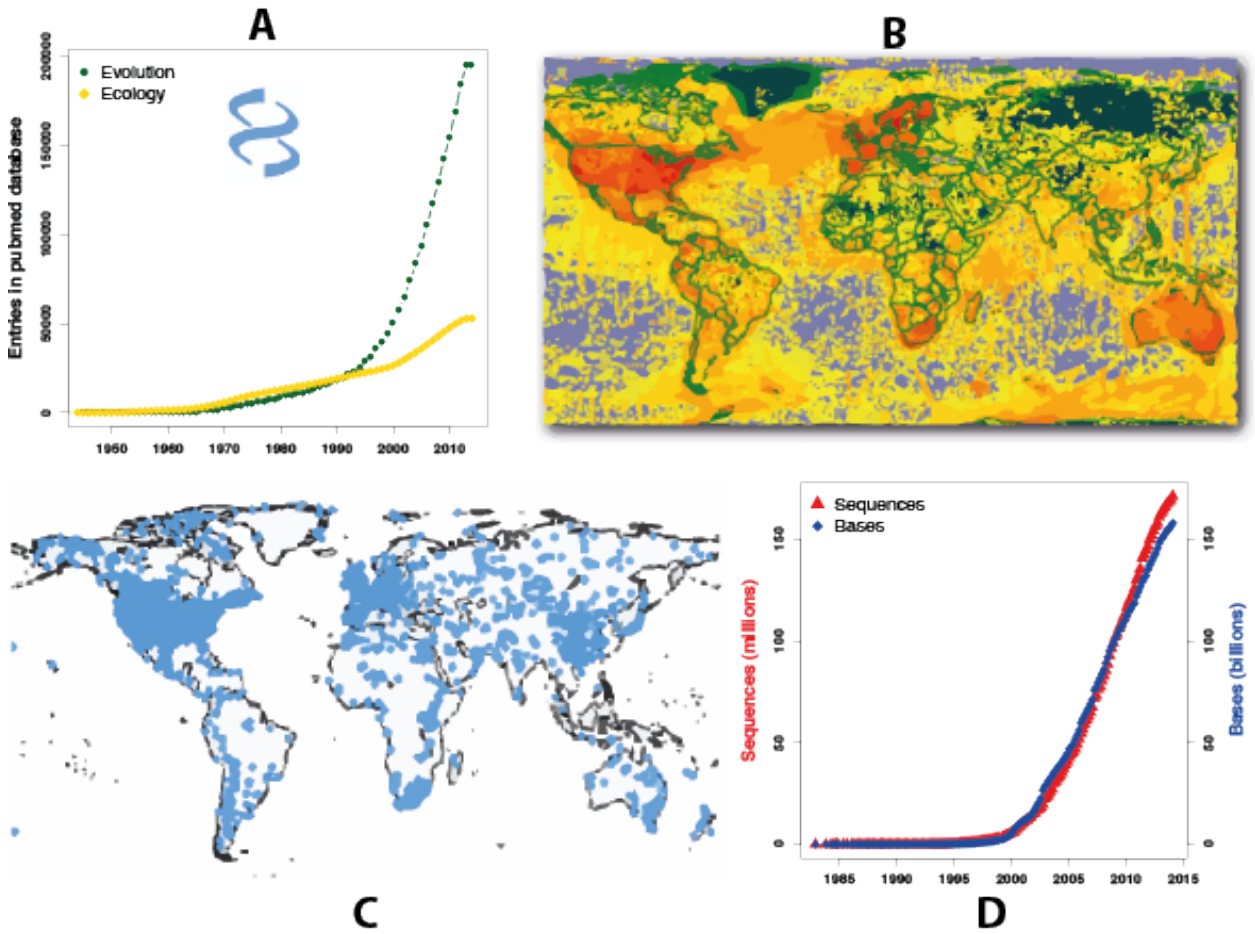
Wiens, J.J. (2006). Missing data and the design of phylogenetic analyses. *J. Biomed. Inf.*, 39, 34-42.
79.

Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M. *et al*. (2007). How Global Is the Global Biodiversity Information Facility? *PLoS ONE*, 2, e1124.
80.

Zanne, A.E., Tank, D.C., Cornwell, W.K., Eastman, J.M., Smith, S.A., FitzJohn, R.G. *et al*. (2014). Three keys to the radiation of angiosperms into freezing environments. *Nature*, 506, 89-92.

1032 **FIGURES**

1033 **Fig 1.** Biodiversity research has entered the era of Big Data. **A:** Growth of the number of
1034 scientific publications in ecology and evolution (search for subject headings in the PubMed
1035 database). **B:** Snapshot of species occurrence records in the Global Biodiversity Information
1036 Facility. **C:** Fossil occurrences available from the Paleobiology Database. **D:** Growth of
1037 molecular data deposited in GenBank.



1038

1039

Fig 2. Methods for inferring large (dated) phylogenies. Schematic comparison of the supertree, supermatrix, and the SUPERSMART approaches (presented in this paper).

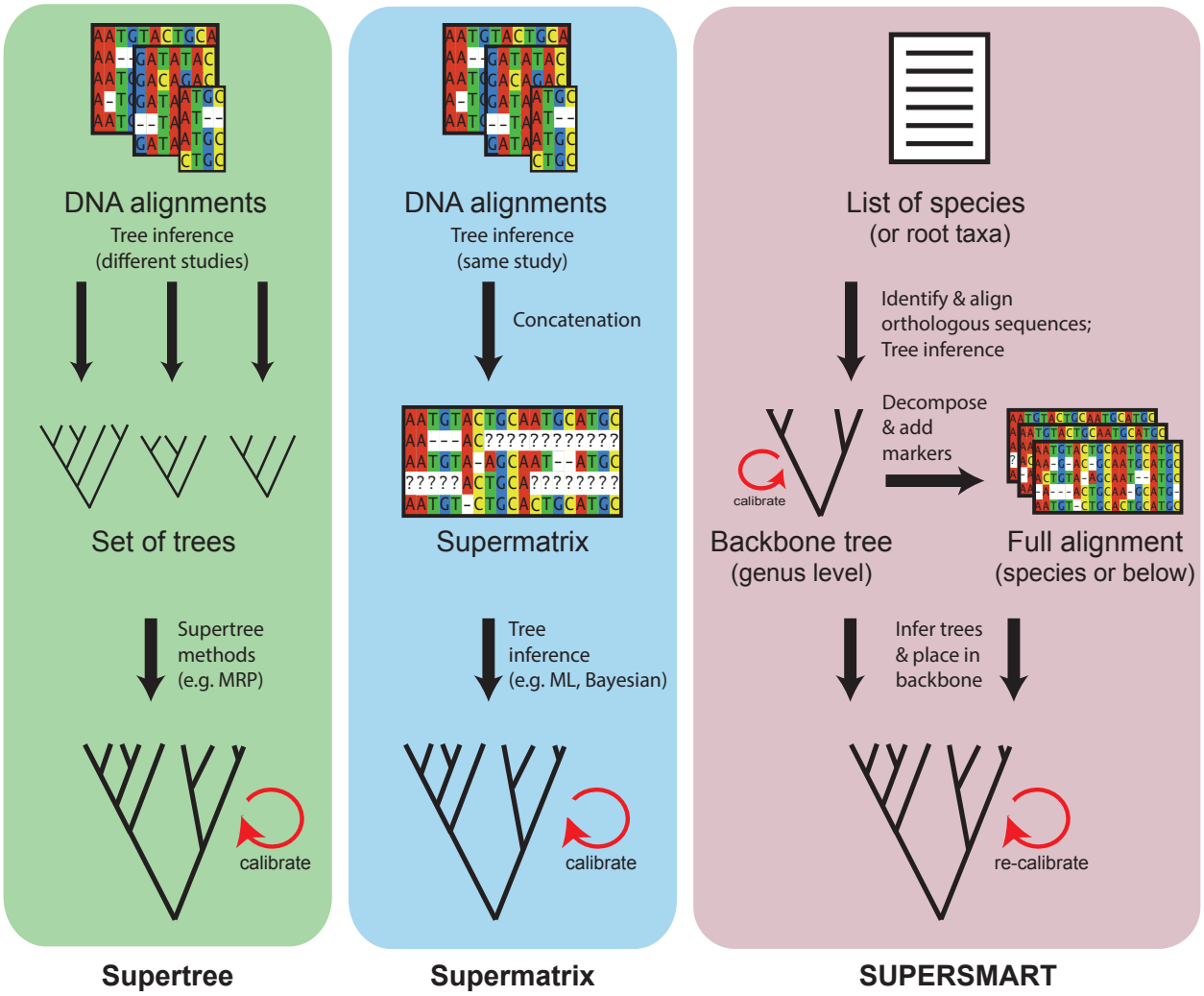


Fig. 3. Basic overview of SUPERSMART. **Step 1**: a hypothetical, dated backbone phylogenetic tree comprising 16 genera, each represented by a single species. The dating employs all available fossils suitable for time calibration. **Step 2** depicts a clade (2 genera with 11 species in total) for which no direct fossil calibration is available. Divergence times can nevertheless be confidently estimated by re-calibrating (scaling) the branches with the posterior clade ages obtained in Step 1. **Step 3:** The fine-level, dated trees are then grafted together and made available for various post-tree analyses in ecology and evolution, relying on directly comparable divergence times.

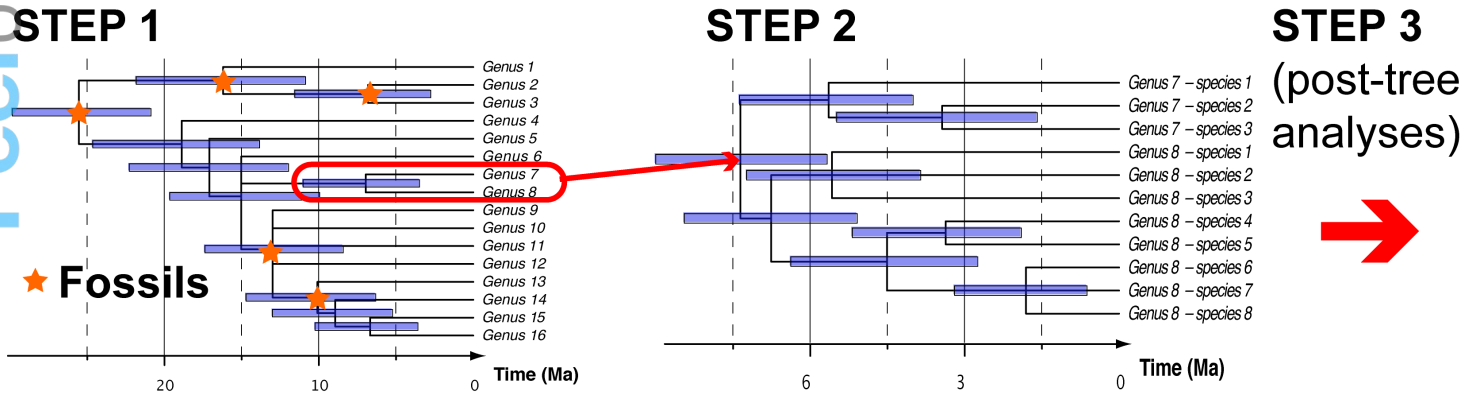
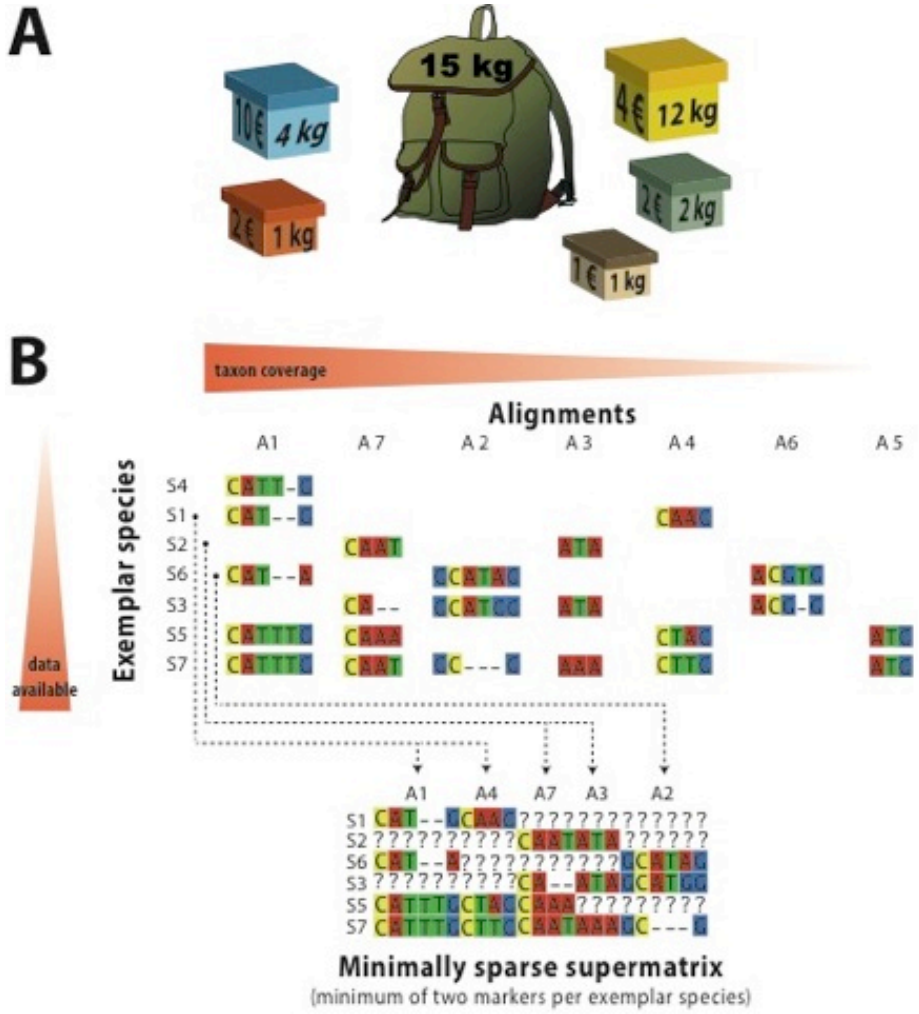
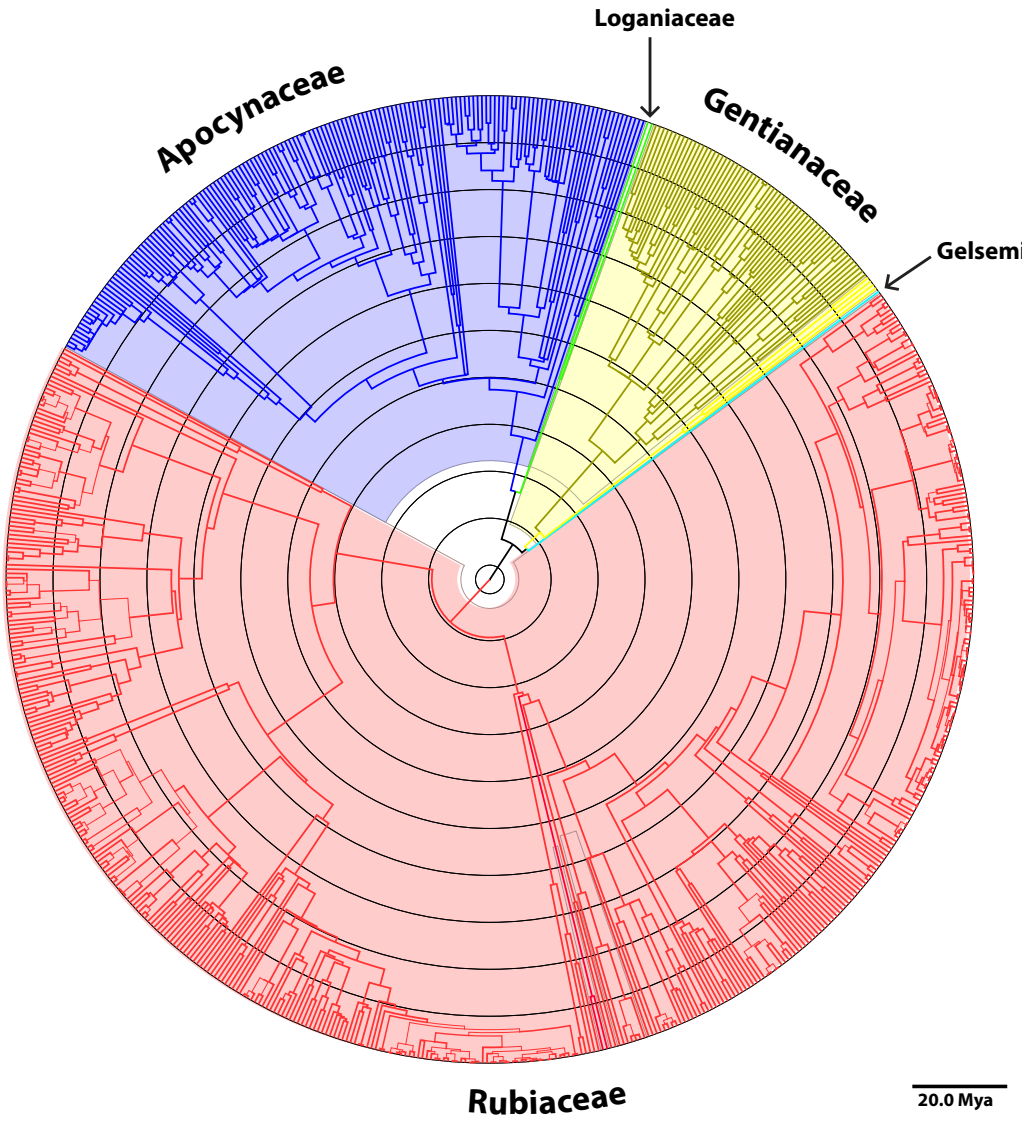


Fig. 4. Illustration of the classic knapsack problem, applied to the optimal choice of species and alignments (markers) for compiling DNA alignments. **A:** Combinatorial problem of maximizing the amount of money in the knapsack with a maximum capacity of 15 kg. **B:** Seven exemplar species (S1-S7) are put in ascending order by their occurrence in the candidate alignments (A1-A7) which are in turn ordered by taxon coverage. In this example, the minimum number of alignments per species is set to two. The supermatrix is then compiled as described in the text. The resulting matrix consists of five alignments and only six species, since the number of alignments in which species S4 occurs does not meet the required minimum.



1067 **Fig. 5.** Time-calibrated phylogeny of the plant order Gentianales constructed using
1068 SUPERSMART, including 701 species and calibrated using 10 fossils. The five families are
1069 outlined. Internal concentric circles represent 10 Ma bins. See Fig. S2 for a fully annotated tree.
1070 The figure was generated with FigTree (http://tree.bio.ed.ac.uk/software/figtree/).



1071
1072
1073

Fig. 6. Molecular chronogram of Primates constructed with SUPERSMART, shown in the tree visualisation tool OneZoom (Rosindell & Harmon 2012; http://www.onezoom.org/). Top left: Full tree; top right: zoom to the genus *Saguinus*; bottom: zoom to a branch comprising four species. An interactive image is available at http://www.supersmart-project.org/p/example.html. See also Fig. S3 for the fully annotated tree in classical style.

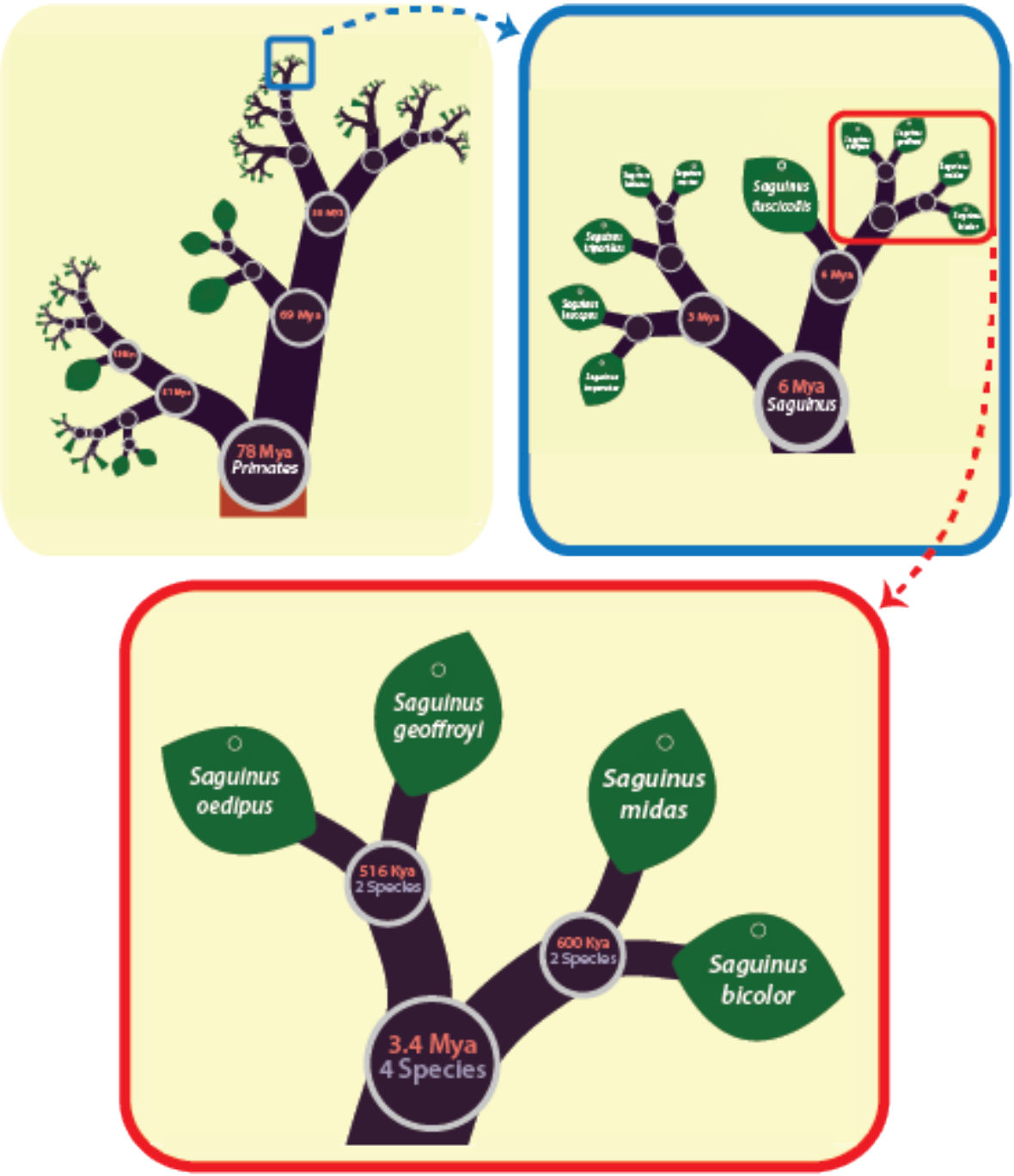
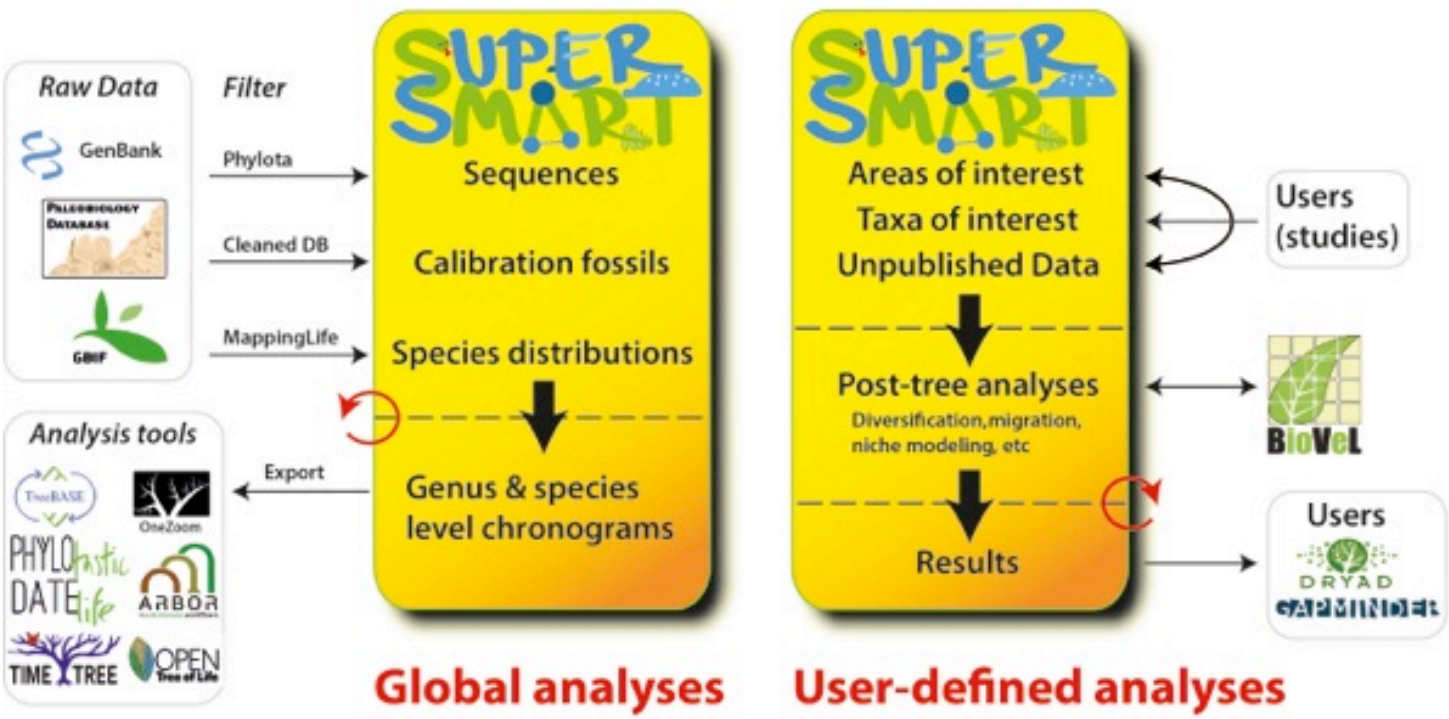


Fig. 7. Proposed interactions with other initiatives at different analytical stages. *"Global analyses"* will provide continuously updating, dated phylogenies of all species with publicly available molecular sequences, and (in upcoming versions) estimates of diversification and migration rates among and within a set of pre-defined GIS polygons (such as WWF's realms and biomes). The results may be retrieved by other initiatives and will be deposited in data repositories. *"User-defined analyses"* are influenced by individual choices, including user-defined polygons (areas), taxa of interest, and fossil records. The user may also include data that are not yet published or are not public. A series of post-tree analyses will be available, mainly through BioVeL workflows, and new ones will be incorporated continuously.



1091 **TABLES**

1092

1093 **Table 1.** Summary statistics for SUPERSMART Steps 1 and 2 of the orders Gentianales and

1094 Primates.

1095

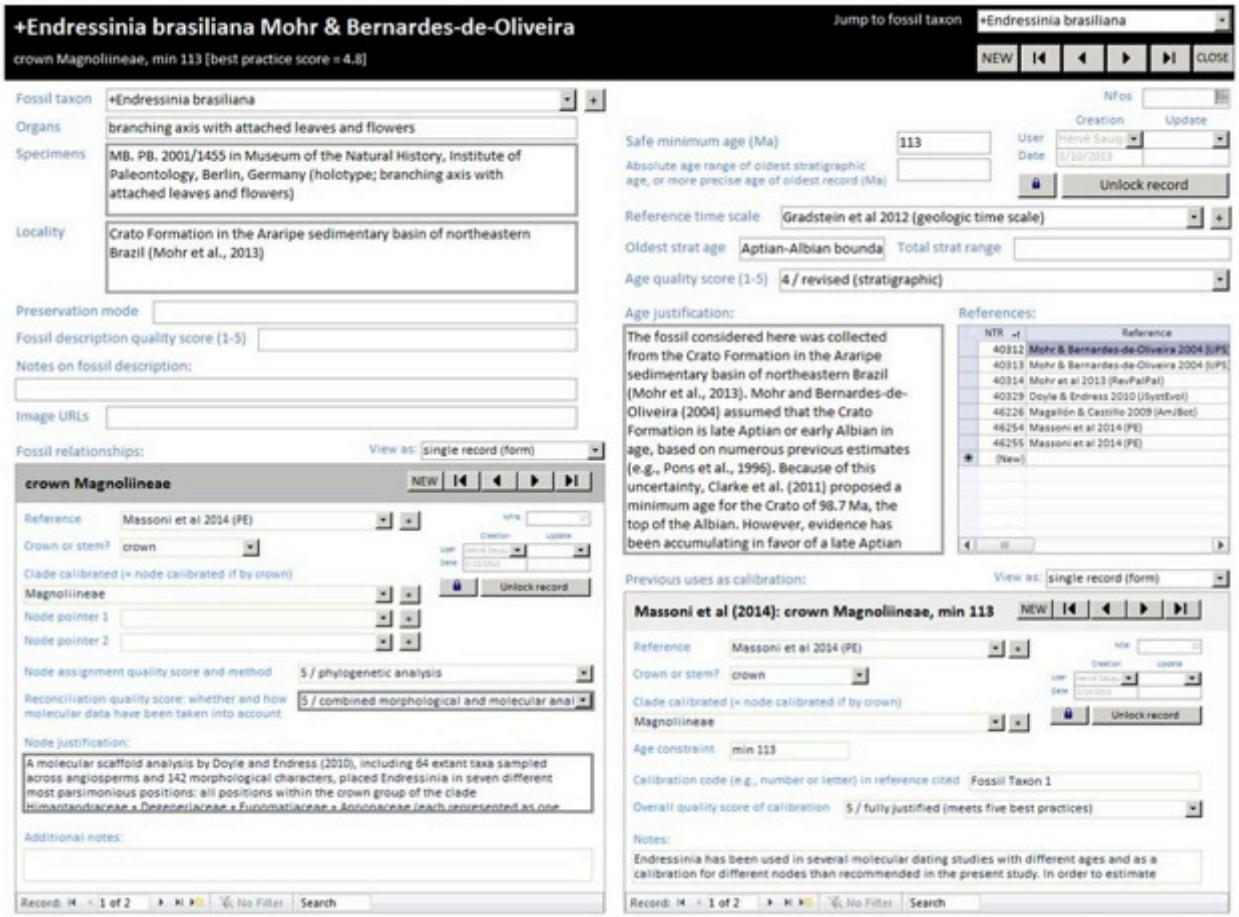
1096

GENTIANALES			STEP 1 (Backbone tree)						STEP 2 (Species-level subclades)				
Families	Genera	Species	Terminals	Loci	base pairs	% gaps	Calibrations	CPU time	Subclades	Total species	Min-Max bp	CPU time
Total (included in SUPERSMART analysis)												
5* 5† (5)	945* 317† (904)	6702* 22237† (5124)	461	18S, 5.8S, 5S, accD, atbB, atpB, atpB-rbcL, atpE, atpF, cox1, CrN1, IGS, ITS1, ITS2, LFY, matK, matR, nad5, nadhF, ndhF, PepC, petB, petN, petN-psbM, phya, progesterone5-beta-reductase, psbA, psba-trnH, psbM, rbcL, RPB2, rpl16, rpoB, rpoC1, rpoC1, rps16, rps3, rps4, tnrC-GCA, tnrC-petN, tnrC-rpoB, tnrE, tnrG, tnrH, tnrK, tnrS-tnrG, tnrY, Tpi, tRNA-Met, tRNA-Ser-GCU, tRNA-Val, trnG, trnK, trnL, trnS-Ycf9	1609382	54	10§	~80h	16	701	593-3611	~60h
PRIMATES			STEP 1 (Backbone tree)						STEP 2 (Species-level subclades)				
Families	Genera	Species	Terminals	Loci	base pairs	% gaps	Calibrations	CPU time	Subclades	Total species	Min-Max bp	CPU time
Total* (included in SUPERSMART analysis)												
16* 14† (16)	73* 60† (67)	494* 249† (201)	93	TrpC2, G6PD, FGA, NPAS3, PNOC, CXCR4, cytochrome b, alpha-2 c2, GPR33, SRY, androgen receptor, 12S, ADORA3, thimidylate synthase, toll-like receptor 5, NEGR1, ZFY, CD44, MC1R, MM1, MM2, St1, TTR, DRB, BDNF, CCR5, irbp, CXCR4	664892	23	8‡	~8h	20	178	93-4288	~60h

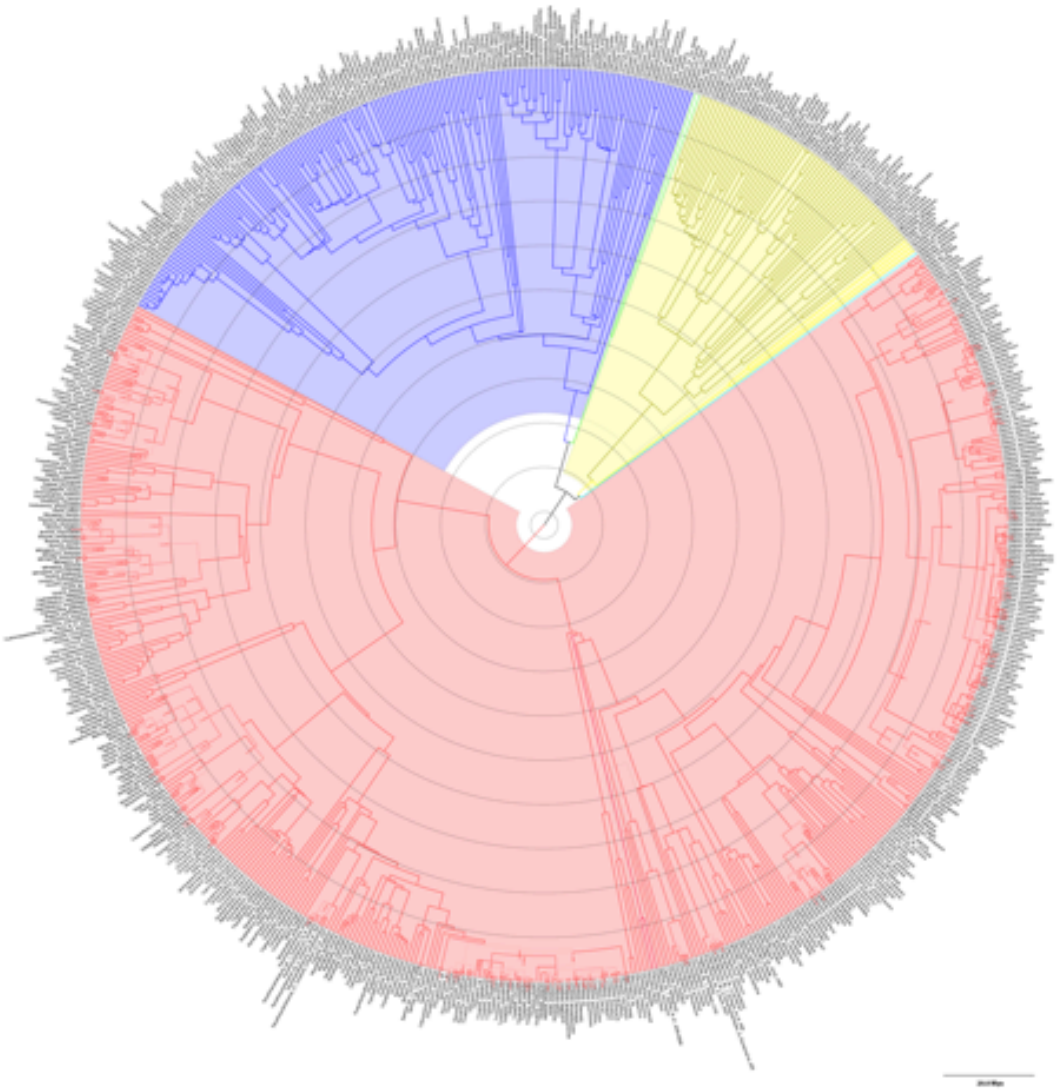
*Statistics from the NCBI taxonomy database (http://www.ncbi.nlm.nih.gov/taxonomy)
† Statistics from the Catalogue of Life (http://www.catalogueoflife.org, June 8, 2014)
‡Vos, 2006

ONLINE SUPPLEMENTARY MATERIAL

Supplementary Fig. S1. Example of a fossil entry in the PROTEUS database. Relevant information such as the taxon name (here *Endressinia brasiliana*), age boundaries, and 'best practice scores' (see text for explanation) can be exported and directly used for time calibration in SUPERSMART.



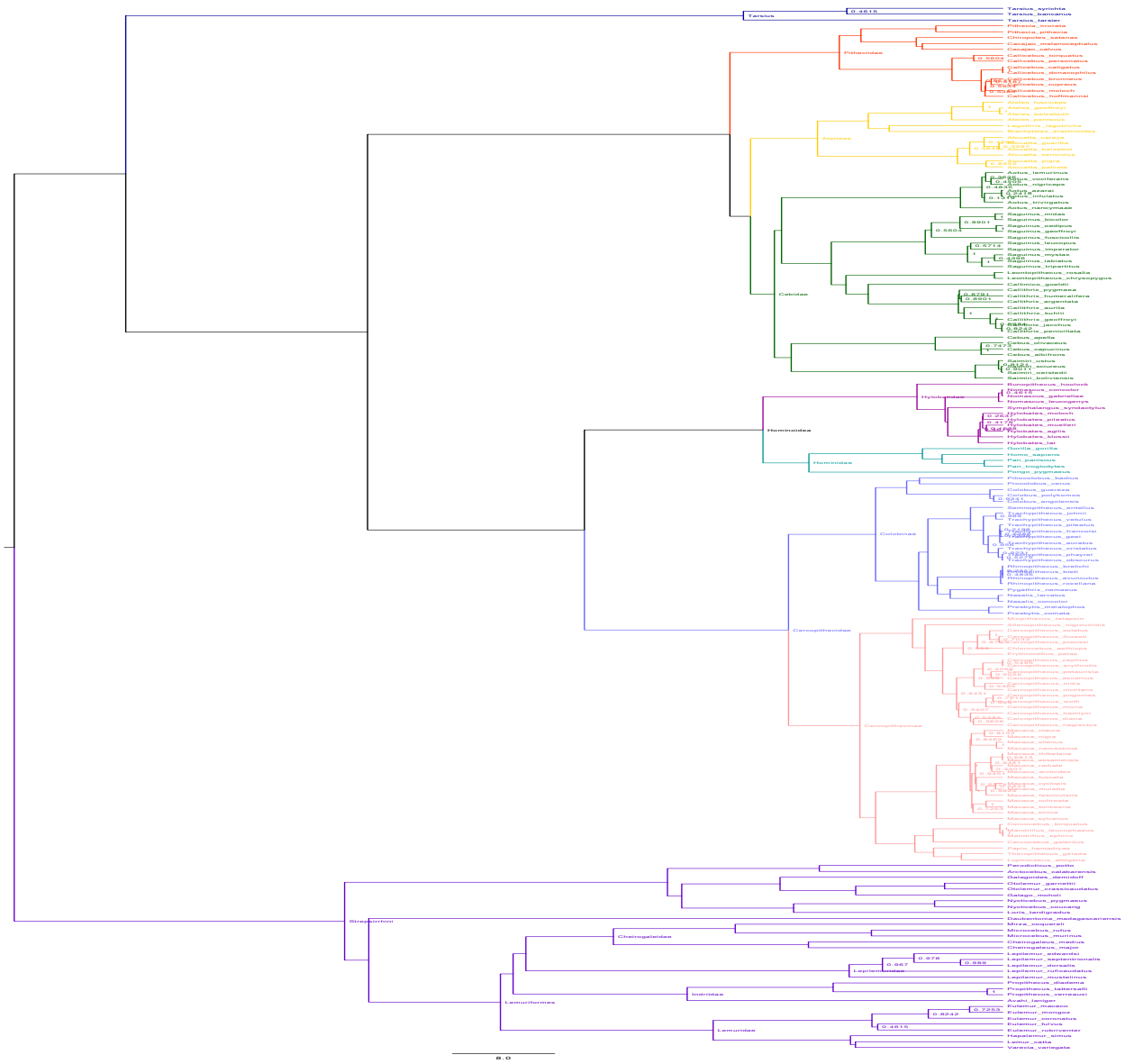
Supplementary Fig. S2. Phylogeny of Gentianales. The final tree inferred using the SUPERSMART pipeline comprises the five families *Apocynaceae* (blue), *Gentinaceae* (yellow), *Rubiaceae* (red), *Loganiaceae* (green), and *Gelsemiaceae* (turquoise). Node labels and branch widths represent posterior support values for inferences of individual clades. The figure was generated with FigTree (http://tree.bio.ed.ac.uk/software/figtree/).



1111 **Supplementary Fig. S3. Fully annotated phylogeny of Primates inferred using**

1112 **SUPERSMART**. Node labels represent posterior support values for inferences of individual

1113 clades. The figure was generated with FigTree.



1114