

TimeTree: A Resource for Timelines, Timetrees, and Divergence Times

Sudhir Kumar,^{1,2,3} Glen Stecher,^{1,2} Michael Suleski,^{1,3} and S. Blair Hedges*,^{1,2,3}

¹Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA

²Department of Biology, Temple University, Philadelphia, PA

³Center for Biodiversity, Temple University, Philadelphia, PA

*Corresponding author: E-mail: sbh@temple.edu

Associate editor: Michael S. Rosenberg

Abstract

Evolutionary information on species divergence times is fundamental to studies of biodiversity, development, and disease. Molecular dating has enhanced our understanding of the temporal patterns of species divergences over the last five decades, and the number of studies is increasing quickly due to an exponential growth in the available collection of molecular sequences from diverse species and large number of genes. Our *TimeTree* resource is a public knowledge-base with the primary focus to make available all species divergence times derived using molecular sequence data to scientists, educators, and the general public in a consistent and accessible format. Here, we report a major expansion of the *TimeTree* resource, which more than triples the number of species (>97,000) and more than triples the number of studies assembled (>3,000). Furthermore, scientists can access not only the divergence time between two species or higher taxa, but also a timetree of a group of species and a timeline that traces a species' evolution through time. The new timetree and timeline visualizations are integrated with display of events on earth and environmental history over geological time, which will lead to broader and better understanding of the interplay of the change in the biosphere with the diversity of species on Earth. The next generation *TimeTree* resource is publicly available online at <http://www.timetree.org>.

Key words: database, software, divergence time, timetree.

Introduction

Divergence times for thousands of individual branches of the tree of life are scattered in the rapidly-growing peer-reviewed literature. To make these results easily accessible to evolutionary and non-evolutionary biologists, we are collecting and synthesizing divergence time estimates from all published studies in the *TimeTree* resource (Hedges et al. 2006; Kumar and Hedges 2011). *TimeTree* is a specialized domain-specific effort, which is needed because the primary information on evolutionary times is published in graphical formats (images) containing hierarchical trees of species. This property of the corpus of scientific results does not permit the discovery of times through usual text mining (e.g., Google text search). Consequently, scientific knowledge on divergence time remained underutilized by a vast majority of scientists and the general public, which is the primary challenge addressed by the *TimeTree* project.

Our initial innovation was to develop a computational tool that utilized the evolutionary hierarchy in capturing greater information than contained in individual studies or their sum. For example, expert and non-expert users wanting to know the time of divergence of two species (e.g., dog and cat) will not easily find an article publishing the split of raccoon and hyena (both trace through the same evolutionary node separating the dog and cat), which provides another source of information for the original query. In fact, studies containing times from different species from the same groups cannot be

retrieved in any simple text search system, irrespective of the degree of their sophistication, because it requires the mapping of the query onto the backbone of the tree of life to discover if the requested species pair traces their ancestry back to a node in the tree of life for which time information is available from other taxonomic comparisons.

Thus, the *TimeTree* project has been advancing the process of knowledge discovery by enabling data-driven retrieval of peer-reviewed publications and estimates of time of divergence between species that are relevant to the user's query (Hedges et al. 2006; Kumar and Hedges 2011). With the availability of the *TimeTree* resource, users have been able to ask questions such as: What is the time of divergence of species (or other taxon) A and B? Is there consensus in the field of molecular evolution on a specific divergence time? What is the range of times that have been proposed? Who are the researchers working on specific divergence times and groups of species? What are their publications? These facilities are being used extensively, which is evident from a large number of *TimeTree* data queries launched (250,000 annually, Google Analytics statistics, 2017 edition) and more than 700 citations to previous publications announcing releases of *TimeTree* (Google Scholar, 2017 edition).

In this article, we describe the next generation of the *TimeTree* resource, which is developed to keep pace with the expanding field (and literature) and more sophisticated

needs of researchers. The literature reporting new timetrees and times of divergence continues to grow at a rapid pace, with hundreds of new studies reporting timetrees each year (fig. 1a), which required us to curate thousands of new timetrees. The complexity of timetrees is also increasing, because the number of species in each study is increasing (fig. 1b and c), which required a modernization of the timetree web architecture. And, we are now able to assemble a large timetree of life based on thousands of studies and species (fig. 1 in Hedges et al. 2015). With this advancement, there was a need to improve web functionality, as users now wish to obtain timetrees of individual clades, including a custom list of taxa, and/or a timeline showing evolutionary branches from the perspective of any selected species, going back in time to the origin of life. The new *TimeTree* resource fulfills these needs, which are important not only for biologists and molecular evolutionists, but also by scientists in other fields such as paleontology and geology, who wish to link the development of the Earth's atmosphere and environment to the evolution of organisms.

Next Generation *TimeTree* Resource

The *TimeTree* resource has a multi-tiered web-based architecture (fig. 2) that includes a non-standard computational

database to store primary data in the form of phylogenetic trees scaled to time needed to compute and serve synthetic data with much greater information content. The new *TimeTree* resource makes a major advance beyond the previous official release in Kumar and Hedges (2011). The number of species is now greater than 97,000 as compared to 17,000 previously, and the number of studies included has increased to greater than 3,000 as compared to 900 previously. The new *TimeTree* resource now adds data from asteroid impacts (Earth Impact Database), solar luminosity (Gough 1981), and atmosphere composition, including oxygen (Holland 2006) and carbon dioxide (Berner 1990; Petit et al. 1999; Hessler et al. 2004; Beerling and Royer 2011). They are all displayed synchronously with divergence times in the form of timepanels. We envision that the co-visualization of the organismal divergences and events on Earth history will lead to new insights into the history of life.

Details about the acquisition of timetree data, their standardization, and the assembly of a global timetree (Timetree of Life, TTOL) based on individual timetrees are available in Hedges et al. (2015), where evolutionary and biodiversity analyses of these data were presented for 50,632 species from 2,274 studies. We added an additional ~1,000 studies and utilized Hedges et al. (2015) method to assemble a timetree of ~97,000 species. In this method, timetrees and divergence data from individual studies are mapped on a conservative guide tree available from the National Center for Biotechnology Information (NCBI) Taxonomy Browser, and these times are used to resolve polytomies and derive nodal times in the TTOL. Hedges et al. (2015) method also evaluates support for each resulting resolution by examining the number of timetree topologies in the database that are concordant with that resolution. When there are more discordances than concordances, alternative topological configurations are tested for each resolution locally and the topology that minimizes the discordances is adopted. It is important to note that many nodes in the tree of life are poorly resolved and the robustness of this solution is limited by existing data. In particular, we recommend researchers to consult the most recent phylogenetic literature, some of which is not incorporated in the *TimeTree* because not all the phylogenetic studies include a timetree.

The new *TimeTree* resource also advances beyond the simple presentation of the divergence time of two organisms (or groups of organisms), which has been the only facility available until now. While this NODETIME search is often sufficient, scientists frequently need more complex information about a species evolutionary history and a tree of species scaled to time. The new *TimeTree* resource provides intuitive ways to access such evolutionary knowledge (fig. 3). In one, the user can now select a set of organisms and view a resulting timetree (the TIMETREE view). This is the most complex view of the data. A simpler view is the TIMELINE view where the user selects a species (or higher taxon), which results in a timeline of ancestral species divergences from the origin of life to the emergence of the selected species. By their basic nature, both the TIMETREE and TIMELINE views involve divergences from multiple time pairs of species and higher level

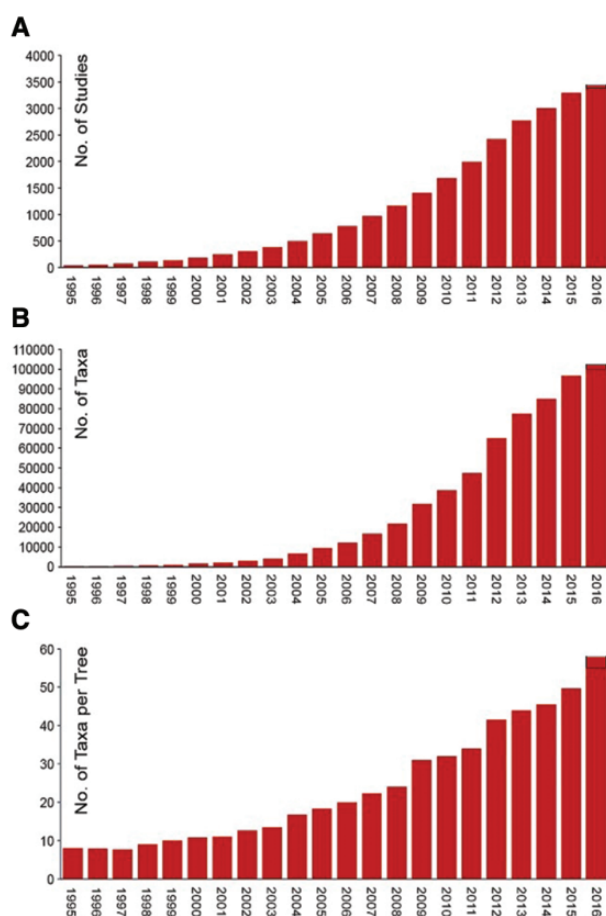


Fig. 1. (a) Number of research publications curated in the *TimeTree* knowledgebase. (b) Number of taxa included in the *TimeTree*. (c) Number of taxa per timetree. All counts are based on cumulative data.

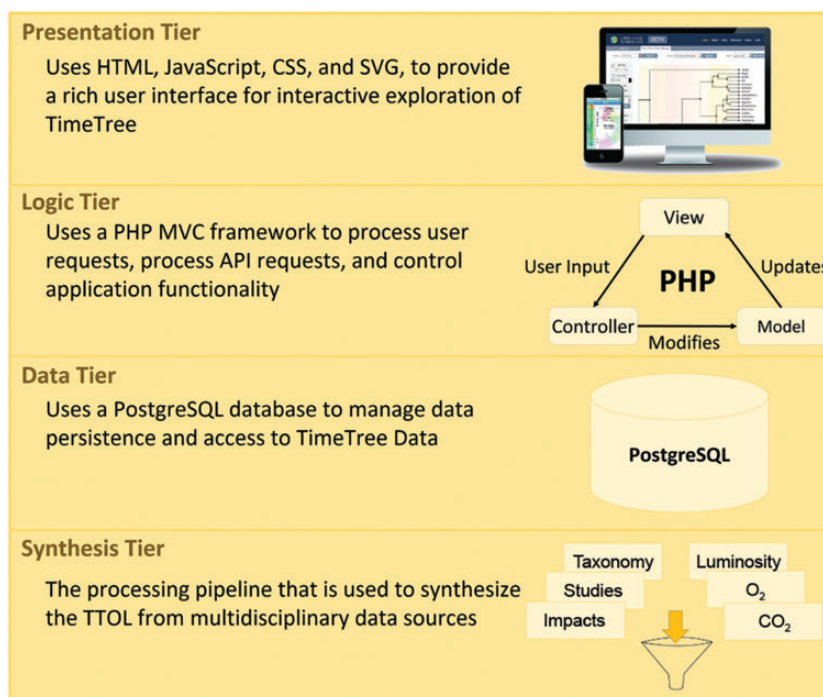


Fig. 2. Four-tiered modular architecture of *TimeTree* that handles the collection and management of scientific knowledge, the translation of knowledge into data, and interactive exploration and visualization. The top-most is the presentation tier, which is the web-based user interface of *TimeTree* and is developed using standard web technologies (HTML, JavaScript, AJAX, and CSS). This tier provides access to all the searches and presents results in SVG vector-based graphics to enable user-interactivity. Directly beneath the presentation tier is the logic tier, which is implemented using a PHP MVC (Model View Controller) framework on an Apache web server. AJAX search requests sent from users' web browsers are handled at this tier by Controllers in the MVC framework and access to data from the backend database is delegated to Models. Search results are returned to the users' web browsers in the form of Views which are composed of HTML, JavaScript, and CSS. The logic tier is connected to the data tier, which manages data persistence and provides access to the information stored in the *TimeTree* database. This is achieved using the PostgreSQL RDBMS (Relational Database Management System) which serves as an API for the logic tier to retrieve data from the *TimeTree* database. The timetrees (which are subtrees of the TTOL) that are produced by the *TimeTree* resource represent a synthesis of multidisciplinary data and, consequently, the data stored at this tier comes from multiple sources, some of which are primary data and others are derived from the primary data. The fourth tier is the synthesis tier, which is responsible for generating the TTOL, and is built semi-automatically using all the source (study) timetrees as input to the Hierarchical Average Linkage methods presented in Hedges et al. (2015).

taxa. In the following, we describe the three primary views in the *TimeTree* resource.

TIMETREE View

This view consists of a timetree with timepanels displayed below (fig. 4). A user can build a timetree in two ways. One, they can request to display a timetree for a given group of species, such as primates. In this case, they will enter the search term (e.g., "Primates") in a text field (fig. 4e) and click the *Search* button. When needed, *TimeTree* will use a synonym list obtained from NCBI Taxonomy Browser to translate the search term from many common names to a scientific name. If the search term resolves to multiple results from the NCBI synonym list, then the user selects one from a drop-down list of possible taxonomic name matches. In the present case, the search term results in only one match. Next, the user restricts the result to their desired taxonomic breadth (e.g., species or families; fig. 3g). In the current case, they can choose to display a tree of 15 families, 77 genera, or 364 species.

Once the user makes the selection (e.g., "Species" in the present case) and clicks the *Show Timetree* button, we display the appropriate timetree extracted from the ~97,000 species

TTOL. To collect the desired timetree, the TTOL is recursively pruned while retaining all the desired nodes and the relevant metadata. This hierarchal timetree is displayed as a rectangular chronogram with nodes marked by circles (fig. 4a). Solid circles mark nodes that map directly to the NCBI Taxonomy and the open circles indicate nodes that were created during the polytomy resolution process described in Hedges et al. (2015). A geologic timescale (showing eons, eras, periods, epochs, and ages) is rendered directly below the timetree (fig. 4b) along with a timescale for divergence times (MYA). Below the geological timescale, timepanels display Earth impacts, solar luminosity, and atmospheric oxygen and carbon dioxide levels (fig. 4c).

Context-Sensitive Information

We have programmed the *TimeTree* server to transmit the timetree display to the user's web browser in the Scalable Vector Graphics (SVG) format, which is an XML-based vector image format with support for interactivity to ensure a dynamic user experience. In the SVG format, we link image elements to metadata and use JavaScript and CSS (Cascading Style Sheets) to add interactive context-sensitive

a **NODE TIME** GET DIVERGENCE TIME FOR A PAIR OF TAXA
Specify 2 Taxon Names
Taxon 1:
Taxon 2:
Clear Search
Resolve Ambiguity
Taxon 1:
Taxon 2:
Show Time

b **TIMELINE** GET AN EVOLUTIONARY TIMELINE FOR A TAXON
Specify a Taxon Name
Taxon:
Clear Search
Resolve Ambiguity
Taxon:
Show Timeline

c **TIMETREE** BUILD A TIMETREE
Specify a Group of Taxa
Group:
Clear Search
Resolve Ambiguity
Group:
Update
Select Taxonomic Rank
Rank:
Show Timetree

Or

d **Load a List of Species**
Upload: Choose File No file chosen
Clear Upload

FIG. 3. The main *TimeTree* Search page, showing (a) NODETIME, (b) TIMELINE, and (c, d) TIMETREE search options. Searches are initiated using common and scientific names which are used to locate scientific identifiers in the NCBI synonym table. Ambiguous names are resolved by selecting possible matches from a dropdown list (f). When building a timetree by specifying a group (c), the user will select a taxonomic rank (g) to define the span of descendants included in the timetree. When using the option to upload a text file with a list of species names (d) a timetree is constructed with the most recent common ancestor of the list of species serving as the root of the tree.

functionality. For example, a user may click on the geologic timescale to get relevant information pertaining to any eons, eras, periods, epochs, and ages (fig. 4e) or click on the Earth impacts panel to get details about a given impact structure (fig. 4f).

A context-sensitive information box is also available for every node in the timetree display (fig. 4d). This box contains taxon name and rank, number of species in that taxon, divergence time, and a hyperlink to the NCBI Taxonomy Browser. The information box also contains buttons labeled *Get Subtree* to separately display a timetree originating with the current taxon and another button labeled *Get Timetable* that when clicked displays a list of studies used to derive the time estimate of the taxon (fig. 4j). In addition, the box contains a confidence interval derived using the Empirical Rule (95%) in statistics, which states that approximately 95% of the times reported across studies will fall within two standard deviations of the mean time when we assume a normal distribution. That is, we use the 95% Empirical Rule for a population of studies, which has an interpretation similar to confidence intervals and hypothesis testing. Intuitively, we expect that the among-study variance derived from study node times as single data points would capture a variety of lower-level errors, including differences in calibration, gene and taxon sampling, and statistical methods used. We present this confidence interval for nodes where times are available from 5 or more studies, otherwise a min–max range of time estimates is given. The Empirical Rule confidence interval is provided as a quick guide to the amount of variation among studies, and we strongly recommend that the researcher review individual studies, their assumptions, and their methodologies

before using any time estimates in downstream research, especially when the precision and accuracy are critical. For this reason, we display a list of all the studies and links to all publications reporting relevant time estimates (fig. 4j).

Timetree Customizations

To the left of the timetree display are toolboxes with options for formatting the TIMETREE view. Users can reduce or increase the spacing between taxon names in the tree, change the tree line color, and select between linear or log scaling of the time axis (fig. 4g). To make it easier to associate times and geological spans, we color the timetree background according to the geological timescale. However, the user can choose the background color palette to correspond to ages, epochs, periods, eras, or eons. Users can also toggle the display of different timepanels (fig. 4h) and save the timetree display in one of several graphics formats (e.g., PNG and JPG) or a phylogenetic format (Newick) (fig. 4i).

Timetrees for Custom Taxon Lists

We have also programmed an option for users to customize the list of species to be included in the timetree. In the *Load a List of Species* option (fig. 3d), a user provides a text file containing a list of species names (one per line). Each name will be resolved to its taxonomic identifier, and the resulting timetree is displayed in the same way as shown in figure 4. On the Results screen, we also show all the taxon names that could not be resolved unambiguously, which allows users to correct or update their input files and try again. When using this option, the TTOL will be pruned to only contain species that are given in the user's list.

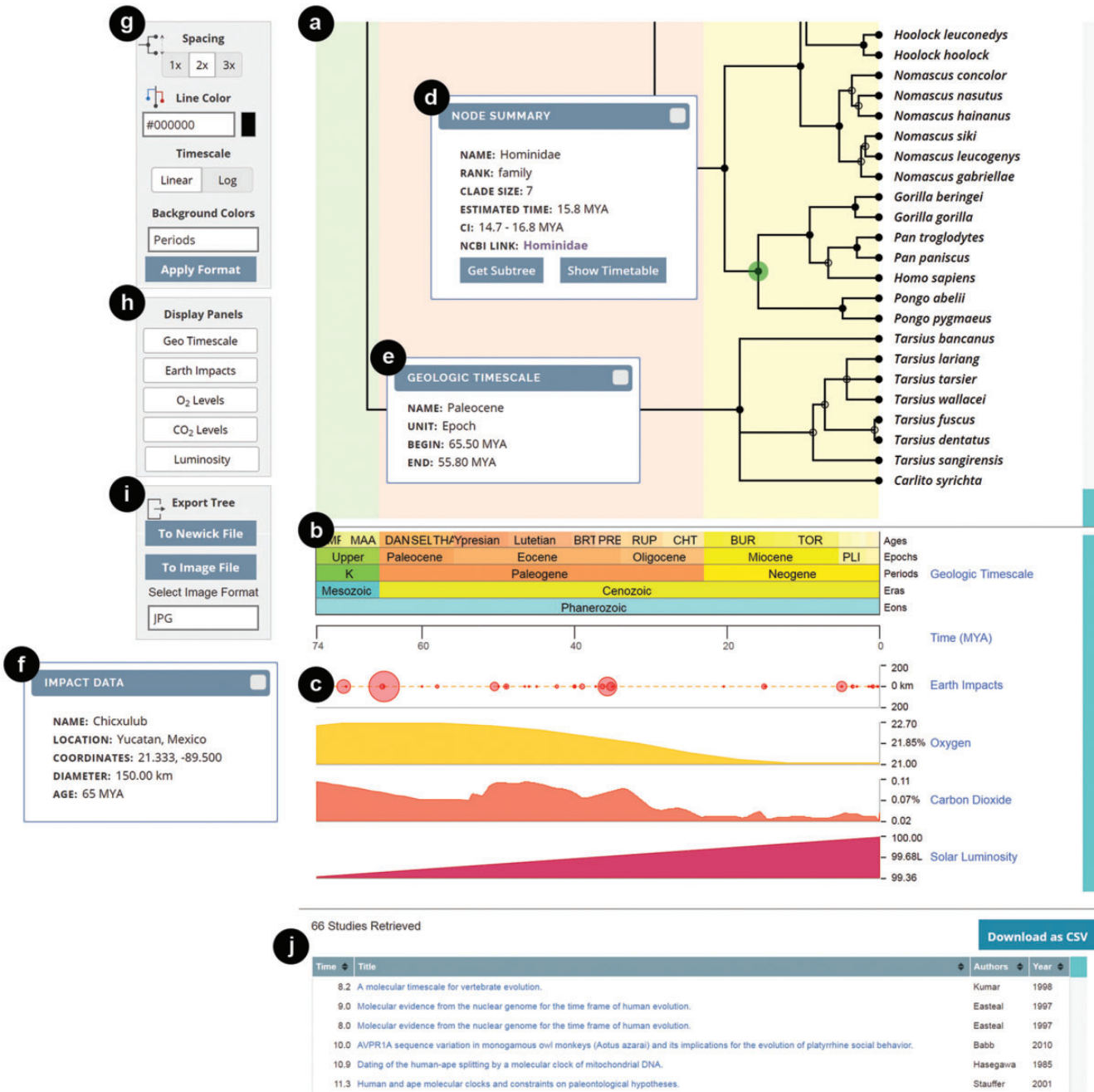


FIG. 4. TIMETREE view showing a timetree for Primates (panel *a*) and featuring timepanels (panels *b* and *c*) and a timetable (panel *j*) of published studies for Hominidae. The SVG images of the tree (panel *a*), geologic timescale (panel *b*), and Earth impacts panel (panel *c*) are annotated with context-sensitive data allowing users to click on specific elements in the images to view detailed information (boxes *d*, *e*, and *f*) or further explore the TTOL. A toolbox on the left provides utilities for toggling the display of panels (*h*), formatting the tree (*g*), and exporting the tree (*i*) to either a Newick file or an image file (e.g., PNG, JPG). Solid circles mark nodes that map directly to the NCBI taxonomy and the open circles mark nodes that are not listed as higher level taxa in the Taxonomy browser, so these nodes are nameless.

TIMELINE View

This view provides the user with a timeline of the evolutionary history for a given taxon (fig. 5). To use the timeline search tool, a user first enters a search term into a text field and then clicks the *Search* button (fig. 3b). Next, if the search term matches multiple queries, the user is prompted to select from possible matches from a drop down list. Once the user makes a selection, they can click the *Show Timeline* button. A timeline is constructed by first locating the queried

species or higher taxon in the TTOL and then traversing the tree toward the root, collecting divergence time estimates at each node visited in the traversal. Solid black circles mark all times along the geological timescale (log-scale) where the taxa have names in the NCBI Taxonomy browser, which are also included in a tabular form. Open circles mark nodes with no corresponding names in the NCBI Taxonomy Browser, and most of these nodes are created as a result of polytomy resolutions. Relevant timepanels, customizations,

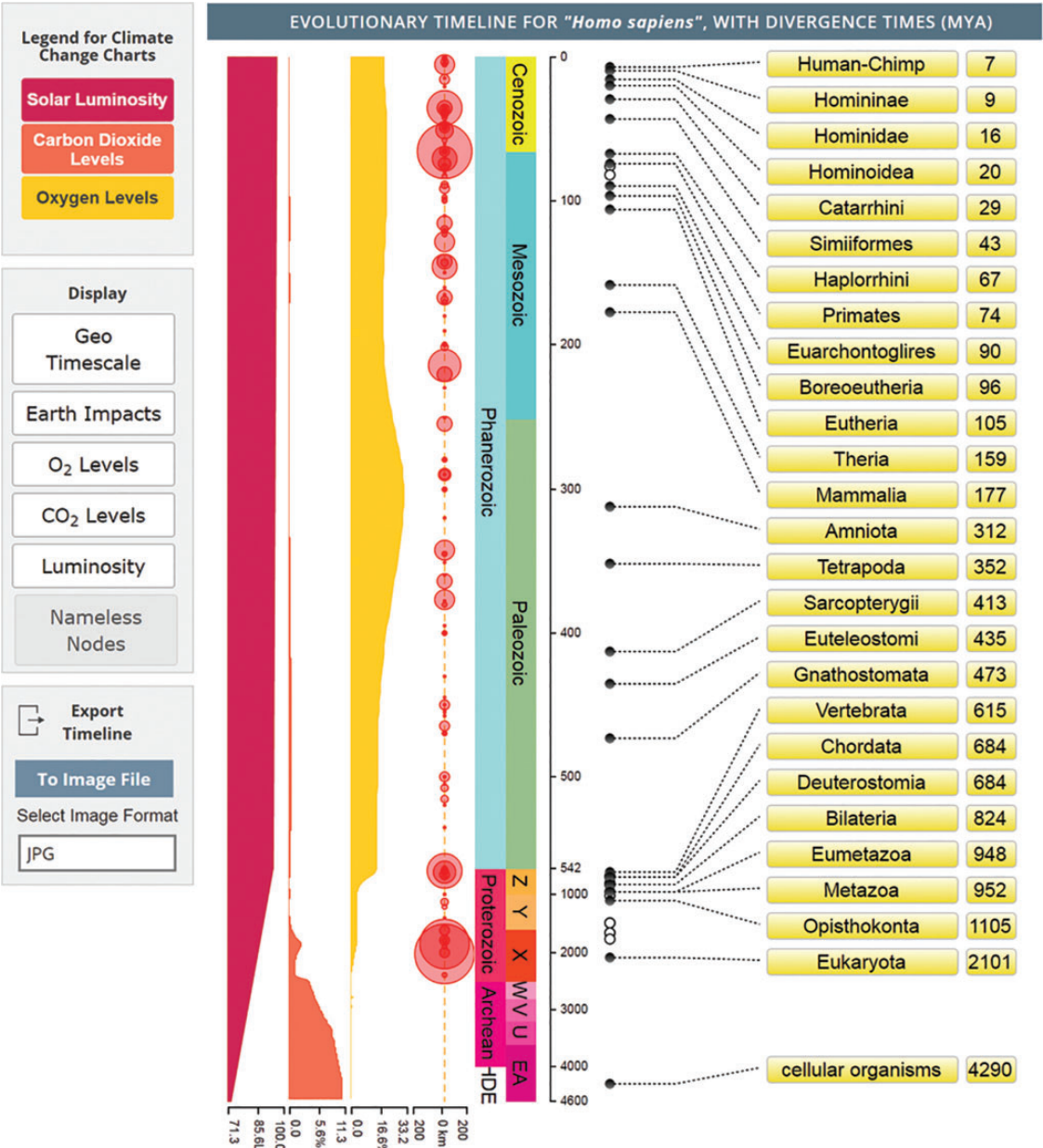


FIG. 5. TIMELINE view showing search results for *Homo sapiens*. The evolutionary history for humans is shown with divergence time estimates at each point along the full lineage. In this view, the geologic timescale and timepanels are rendered vertically and can be clicked to produce the same context information boxes as the TIMETREE view. Solid circles mark nodes that map directly to the NCBI Taxonomy with names and the open circles mark nodes that are not listed as higher level taxa in the Taxonomy Browser, so these nodes are nameless. Display of nameless nodes can be toggled.

and context-sensitive information mentioned above for the TIMETREE view are available for TIMELINE view as well.

NODETIME View

In this view, users provide two names (common or scientific) as input (fig. 3a), which are then disambiguated in order to select proper taxonomic identifiers. Once the query organisms are correctly specified, we first look for both of these species in the TTOL. If they are found, then all studies (and times) that inform about the divergence for the evolutionary split between the two query organisms are presented in the form of a time table (fig. 6). On some occasions, the TTOL

does not contain one or both of the organisms. In this case, we scan the NCBI Taxonomy to locate the most recent common ancestor of the query organisms that is also found in the TTOL. If this is possible, then a time table can be constructed. This new approach is very quick for mining all relevant studies, because the previous method to scan each study individually (Hedges et al. 2006) becomes time consuming due to a large number of taxa in timetrees of each study and the number of studies in the database.

In all cases, we present the median time as well as the divergence time from the TTOL for the organisms of interest. Times from individual studies are marked with a dot along a geological timescale to show the diversity of time estimates

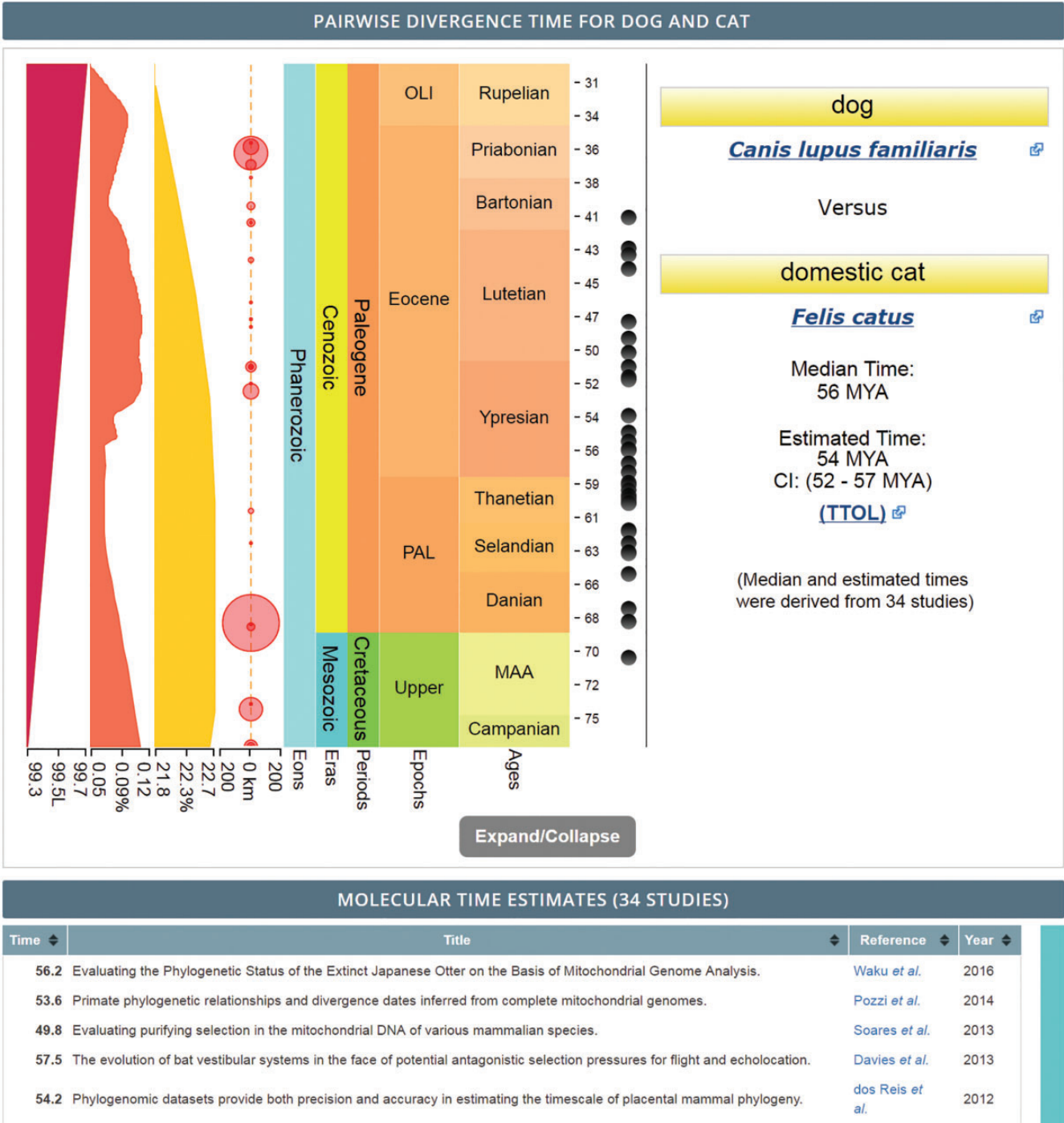


Fig. 6. NODETIME view with pairwise divergence time estimates for cat (*Felis catus*) and dog (*Canis lupus*), along with a timetable of 34 published studies that were used to estimate the time of divergence (5 shown). Published time estimates are marked in the summary view by black circles next to the geologic timescale.

that have been reported in the literature (fig. 6, top panel), and a timetable below the result summary shows detailed information for individual published studies available (fig. 6, lower panel). Links are available for the user to click on taxon names and be taken to NCBI's Taxonomy Browser. All the context-sensitive information and customizations for the timepanels are available for NODETIME view as well.

In summary, the *TimeTree* resource has undergone a major transformation in its presentation of timetrees and has

advanced significantly in terms of the data size (number of timetrees and species). We anticipate adding hundreds of new studies each year to expand TTOL continuously to serve the community needs.

Acknowledgements

We thank K. Boccia, J. R. Dave, K. Davis, A. K. Gawlinski, S. L. Hanson, A. Hippenstiel, M. Li, A. Lu, M. McCutchan, Y.

Plavnik, A. Shoffner, and L. Wu for technical assistance; and authors of studies for contributing timetrees. This research was supported by grants from US National Science Foundation (DBI 1356548 and DBI 1455762) and National Aeronautics and Space Administration (NNX16AJ30G and NNA09DA76A) to SK and SBH, respectively.

References

- Beerling DJ, Royer DL. 2011. Convergent cenozoic CO₂ history. *Nat Geosci.* 4:418–420.
- Berner RA. 1990. Atmospheric carbon dioxide levels over phanerozoic time. *Science* 249:1382–1386.
- Earth Impact Database. Available from: <http://www.passc.net/EarthImpactDatabase/>.
- Gough D. 1981. Solar interior structure and luminosity variations. *Sol Phys.* 74:21–34.
- Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* 22:2971–2972.
- Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. 2015. Tree of life reveals clock-like speciation and diversification. *Mol Biol Evol.* 32:835–845.
- Hessler AM, Lowe DR, Jones RL, Bird DK. 2004. A lower limit for atmospheric carbon dioxide levels 3.2 billion years ago. *Nature* 428:736–738.
- Holland HD. 2006. The oxygenation of the atmosphere and oceans. *Philos Trans R Soc Lond B Biol Sci.* 361:903–915.
- Kumar S, Hedges SB. 2011. TimeTree2: species divergence times on the iPhone. *Bioinformatics* 27:2023–2024.
- National Center for Biotechnology Information (NCBI) Taxonomy Browser. Available from: www.ncbi.nlm.nih.gov.
- Petit J-R, Jouzel J, Raynaud D, Barkov NI, Barnola J-M, Basile I, Bender M, Chappellaz J, Davis M, Delaygue G, et al. 1999. Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* 399:429–436.