Participants: Brian O’Meara, Luke Harmon, Jonathan Eastman, Peter Midford, Tracy Heath, Joseph Brown, Matt Pennell, Mike Alfaro

Day 5:

- To do: work on getting additional fossil calibrations in our local FDB (Tracy)
- To do: Decorate input tree with fossil calibrations (Tracy, Brian)
- To do: NeXML parser/writer to work with phylo objects (Joseph, Peter)
- To do: Add Congruifier to datelife.org (Brian)
- To do: Other output formats (newick, R object, etc.) to datelife.org (Brian)
- To do: Load more PhyloOrchard trees, including ones in the embargo branch. (Brian)
- To do: interface with TNRS
- To do: discuss near-term goals
    - DateLife paper, some thinks to think about:
        - General outline
        - Examples/use-cases
        - Target journal
        - Author order (Brian takes lead)
    - lightning talk on datelife at iEvoBio (it’s probably best to submit the abstract for this soon, since there might be a lot of people who will try to do lightning talks since Evolution is full up)
-

Day 4:

- Create a list of things that will make up the metadata returned by our products [done]. This will help us work toward developing a vocabulary for NexML for fossil calibrations, warning and other stuff specific to our tools, citations, etc [done, below].
- create an example NeXML file with meta data using the vocabulary [done]
- Expecting input to be NeXML. Still unclear if we are getting that from topology or from treestore.
- Vertebrate tree now contains 15,104 tips. Working on improving samples for both snakes and amphibians and might be able to get close to 20k.
- Incorporated the vertebrate tree compiled by Luke et al into our demos, and trees from the 10Ktrees. They are all now in phyloOrchard [done], (we can add some fossil calibrations that will be applied to this tree, but not sure it’s worth it today)
    - cleaned up tip labels (some had accession numbers, etc.)
    - ‘phylo’ format used for each tree accession
    - each tree has citation and clade name (e.g., Mammalia, Spermatophyta)
- DateLife now can summarize the ages of nodes across a cloud of trees and displays the results in a basic table and shows the results from any trees in PhylOrchard that contain the query node: see example
- To do: smart tree-grafting (glomograms)
    - resolve internal node labels via TNRS for a given skeleton tree

    - match internal labels with available subtrees in tree store and glom together
- To do: Webify Congruifier
- To do: allow for queries to the temporary fossil calibration database, discuss output [done], and user input

NeXML Vocabulary Information

Meta-data associated with the scaling-services tree for NeXML and vocabulary (this is being hosted by datelife.org):

- We can make switches that can be thrown by the user to turn on/off some of these metas, as well as allow export as newick or other formats
- For the whole file:
    - **analysisDate**: Date this was done
- For the tree as a whole
    - **branchLengthService**: Description of the service that got the branch lengths (datelife or Congruifier), this will include the version number
    - **serviceOptions:** Description of options that were used
    - **[deprecated] serviceVersionNumber**: DateLife or Congruify version number (there will be bugs and improvements, this allows for reproducibility)
    - **branchLengthSourceTreeCitation**: A citation of tree(s) from where the branch lengths were obtained (e.g. cite of the B-E mammal tree paper)
    - **userTreeCitation**: A citation of the tree we're dating (could be user, but could be from phylotastic)
        - should indicate dating method (PATHd8, r8s, BEAST, etc.)
    - **serviceErrors**: A list of errors/problems if any
    - **treeSet**: If the tree is associated with other trees (e.g. an MCMC set of trees or multiple trees from different r8s runs) there needs to be something to indicate that and link it to other trees; or if the tree is a summary of multiple trees
- For the tips:
    - Much of the data associated with the tips will come from other services (probably? TNRS?), but there may be metadata we find useful to add to tips
- For the edges:
    - **length**: branch length in proportion to time (NeXML already has a syntax for branch lengths)
- For the nodes:
    - **nodeAge**: Age
    - **ageSummary**: If it's a summary of multiple trees, there could be summary statistics of the age (these are nested within the ageSummary identifier)
        - **ageRangeSummary**: 95% Credible interval or high/low age
        - **medianAgeSummary**: median age
        - **meanAgeSummary**: mean age
    - **nodeSupportValue**: Support value

    - **taxinomicName**: Node name, if any, from taxonomy (e.g. Glires)
    - **nodeSpecificErrors**: Per node error or cautionary messages (such as the node being made a polytomy due to dating conflicts)
- For calibrated nodes:
    - **fossilSpecimenName**: Fossil specimen name
    - **calibrationSource**: Name of the source of the calibration (biogeographical or secondary calibration)
    - **calibrationSummaryAge**: “Best guess” age: perhaps the mean or median age (mean makes sense for a single Bayesian run, but when aggregating data across many analyses, one really bad study could affect mean more than median)
    - **minimumNodeAge**: Minimum age for the node from the fossil calibration (this may be a range of dates)
    - **maximumNodeAge**: Maximum age of the node, if available (if there is a maximum age, there might be a different citation for this)
    - **ageProbabilityDistribution**: Probability distribution for the age (perhaps in BEAST format, or a standard ontology for log normal, normal, etc.)
    - **fcdbEntry**: A URL that links to the entry on FCDB (or for the moment the url to Date-A-Clade)
    - **fcdbCitation**: If the fossil is from FCDB, then a link to citation for the FCDB
    - **pbdbEntry**: link to the PaleoBiology DB (PBDB) entry for the fossil (if available, though, this might not be entirely necessary if the calibration comes from the FCDB)
    - **fossilCitation**: Citations associated with the fossil (e.g. the citation describing the fossil, this might not be available from calibration database)
    - **validationCitation**: If the fossil is from the Fossil Calibration DataBase (FCDB), then there will a citation validating it as a fossil calibration (most likely published in Paleontologica Electronica)
- Possible for the future?:
    - If the trees are from BEAST (or other divergence time software like DPPDiv or RevBayes), then they can have other branch/node-associated parameters like substitution rate, speciation/extinction rate, fossilization rate, etc. And if the tree is a summary tree from MCMC output of those programs, then there will also be summary statistics of those parameters (e.g. 95% CI, mean, median). Also make sure to allow any metadata sent in to roundtrip back out
    - Perhaps a description of the method that generated the times of the tree: NPRS, PL, Bayesian inference under an auto-correlated log-normal model (though this could get to be a lot of information)
    -

Day 3:

- Work done on http://datelife.org : it can now give you age, warning message, and even a chronogram (if you give it three or more taxa). Waiting on a tree store for more trees and

TNRS for name resolution, though we may use a stand-in in interim (NCBI)

- Will also roll congruifier into this resource
- We can now deal with a cloud of trees from one study efficiently
- Coordinated with fossil calibration database group. Their product will be useful but hidden for at least the next six months
- Work done on temporary storage of fossil ages
- Work done on R <-> NExML import (<- not yet!) and export (<- works)
- Need to store and present citation info for studies/fossils and include the metadata describing the method used to add branch times or fossil calibrations (e.g. datelife or congruifier), as well as any issues
- Still need to discuss export of calibrations/priors for Beast, r8s, etc. Debate about how much to require of users (e.g. we might want to provide pre-formatted calibrations/priors for teaching purposes)

Day 2 Plan:

- Temporary (?) datastore: PhyloOrchard or private repo of unpublished chronograms (**Luke** and **Mike**)

Two approaches to getting dates:

- Congruifier (**Jon**): a method for scaling trees to units of time
    - Read 'stock' tree from tree store or from web service
        - stock tree has been dated with actual fossil calibrations
        - branch lengths in units of time (ultrametric)
        - this tree will donate node heights
    - Read 'scion' tree from tree store or from web service
        - scion tree has molecular branch lengths (in units of change/time)
        - this is the tree that requires time-scaling
    - Read taxonomic reference table from datastore (or do on-the-fly lookup)
        - still need to get the exact format of the TRNS reference table
        - currently 'congruifier' uses a hash key system for identifying nodes in order to match nodes from 'stock' to 'scion'
        - this method is fairly efficient and does not rely on having proper names for matching nodes (but instead relies on the unique hash keys)
    - Reconcile tips in the 'stock' and those in the 'scion', identifying concordant nodes
    - Return table of reconciled nodes (i.e., 'congruification table')
        - spanning taxa to uniquely identify node in 'scion'
        - node height
    - Allow for immediate time-scaling based on 'congruification table'
        - pathd8 (requires **no** user input)
        - r8s (demands user input)
        - treePL (demands user input)
        - phylocom bladj

    - Action items our group
        - handle fossil constraints (in addition to or instead of ‘stock’ tree)
            - interact with NESCent working group on fossil calibration - in progress (via Daniel Ksepka @ NCSU).
        - code to get user input

- DateLife: Like TimeTree, but open, reusable, uses sets of trees from a single paper to return uncertainty, can take topology and do dates

Note that *since this morning*, we have created code that can use the 4,510-taxon Bininda-Emonds et al. mammal tree (and presumably any other chronogram) and for any set of taxa in the tree give the age of their MRCA. We also have registered a domain name and have a stub at that site (http://datelife.org)

The purpose of DateLife is to obtain a TMRCA for a given set of taxa

    - Read trees from datastore
    - Convert to a usable R format (**Joseph** on NExML -> phylo)
        - first concentrating on phylo -> NEXML (done)
    - Convert to patristic distances (distance in terms of branch lengths between two taxa)
    - Cache patristic distances
    - For a user query, filter for the relevant entries in the matrices (first filtering for those matrices with the relevant post-TNRS taxa), find the max distance.
    - Given ultrametric trees with brlen in units of MY, half the patristic distance = age of MRCA
    - Return this to user. Cam’s interested in RDF return of the date and other metadata such as what genes were used to generated the phylogeny
    - Allow query with more than two taxa
        - If user selects “give all dates”, return ages of all pairs of taxa
        - If user selects “crown group age”, and all taxa are in the database, return the age as you do for two species
        - If user selects “crown group age”, and a subset of taxa are in the database (i.e., ask for all great ape species, only have chimp and bonobo), have option to
            - Return missing
            - Return best guess with a warning that it’s an underestimate of age
        - If user selects “crown group age” but only includes a subset of the taxa in the clade, have option to
            - Return “missing”
            - Return best guess with a warning that it’s minimum
    - Allow query with topology (used by phylotastic)

        - Get ages of all nodes by using multi-taxon queries like above (for each node, get list of descendant taxa, get age as above)
    - Allow specification of higher taxa: get age of MRCA of “angiosperms” and “pines”
        - Requires phyloreferencing and/or TNRS
        - Not sure at this point whether we can use common names as an alternative to proper names (this gets into phylocode debate which is a huge can of worms)
    - Action items our group
        - Code to do the above (**Brian**)
        - Code to get user input (web interface)
        - Code to present output
        - Documentation and examples
        - ~~Buy datelife.org domain name~~ [done!]
    - Action items for other groups
        - Get specification of RDF trees (**Peter**)
        - Figure out how sets of trees are lumped (i.e., all MCMC samples from a Beast run should be linked, so we can say, “This study has a mean age of 15 MY, with a 95% credible interval of 10-17 MY”)
        - We would like people to be able to input a set of trees from a MrBayes/BEAST run or a collection of bootstrapped trees which are dated with PL or whatever
        - Higher level names
        - What output do you want??
            - NExML (and if people want RDF, convert using some downstream tool)
    - Cool things for future or people not directly working on above things (**Tracy**, **Matt. Joseph**, talk to Karen about fossil group at NESCent)
        - This provides for an important a use-case for Phylotastic that gives the user a starting tree and list of phylogenetically-placed calibrations for application in a divergence time analysis.
        - Store fossil dates rather than just chronogram dates - looking into possible existing database generated via previous NESCent working group
        - Way to assign fossils to nodes and add/return metadata
        - **UPDATE:** I’ve (Joseph) talked to Daniel Ksepka, and he is interested in discussing this. They are presently in the building stage of their database now, so things may not be available for a while. Daniel is at NCSU, and so can potentially drop into NESCent if people are willing and able to work with him. (Joseph, can you share the email with the group? It’d be great to contact Daniel and invite him to come work with us or at the very least de-brief us on how we might be able to interface with their db in the future...) Email sent.

- This is happening 6 June at 13:30 NESCent time.

Day 1 Notes:

To do:

Web interface to congruifier

Tree store of chronograms, esp. sets of post beast/ post r8s bootstrap trees (i.e., not just point estimates). Think of what metadata Karen’s group needs to store to properly scare users

Way to pull dates from the tree store for a pair of taxa (need TNRS to identify the taxa)

Way to serve these dates to a user

Way to make up missing edge lengths and communicate that to users (bladej, birth death, uniform, not giving any back, giving many different ones back)

What other caveats, EULAs for users

Users supply:

- Pair of taxa (get back range of ages)
- Topology (get back chronogram(s))
- Tree with molecular brlen (get back chronogram (or many with uncertainty))

Concrete infrastructure needed:

- Chronogram store, with metadata
- TNRS interface
- Web <-> R interface for congruifier
- congruifier -> chronogram interface
- for a given tree, for a pair of taxa, return age MRCA
- summarize MRCA across relevant trees

Samples:

the main function in the package for our purposes is 'congruify.phylo()' -- you can run an example of it with 'example(congruify.phylo)' at the R console

The most recent package is bundled into auteur_0.666.0525.tar.gz -- can be installed with 'R CMD install auteur_0.666.0525.tar.gz'

Code:

auteurDev on bitbucket

Notes:

we should probably assume the trees will be nexml. In which case, we'd need a nexml -> R translator

- test cases here: https://www.nescent.org/wg_evoinfo/NeXML_Test_Files
- apparently this is in the works for NCL:
  - http://informatics.nescent.org/wiki/Phyloinformatics_Summer_of_Code_2012#Extend_NCL_.28the_NEXUS_Class_Library.29_to_parse_nexml.2C_phyloxml.2C_and_the_NOTES_block_in_NEXUS
  - http://informatics.nescent.org/wiki/PhyloSoC:Extend_the_Nexus_Class_Library_to_parse_NeXML_and_PhyloXML
  - as phylobase uses NCL, this is done if NCL is updated

JWB - got some rudimentary phylo (i.e. APE) -> NEXML R code code (shared in Phylotastic folder as "NEMLer.R"). horrifyingly ugly, but works for rooted or unrooted trees with float edge lengths. have not yet considered case where 1) root edge exists, 2) no edge lengths exist. also have not dealt with meta data, but seems straightforward (once I know what meta data will be involved). may work on this to build a real xml object; may facilitate eventual NEXML -> phylo R code.