



# **Self-Updating Platform for Estimating Rates of Speciation and Migration, Ages and Relationships of Taxa (SUPERSMART)**

Rutger Vos & Karin Nilsson

**Naturalis**  
Biodiversity  
Center



# Goals

To develop an analysis workflow that:

- produces supermatrices of aligned sequences for a set of taxa of interest
- implements clever divide-and-conquer methods to build trees of any size from these supermatrices
- makes the resulting trees amenable to re-calibration using fossil anchors
- integrates the resulting chronograms into phylogeographical analyses
- makes these trees consumable by other phyloinformatic projects, such as OpenTOL, PhyloTastic and BioVeL



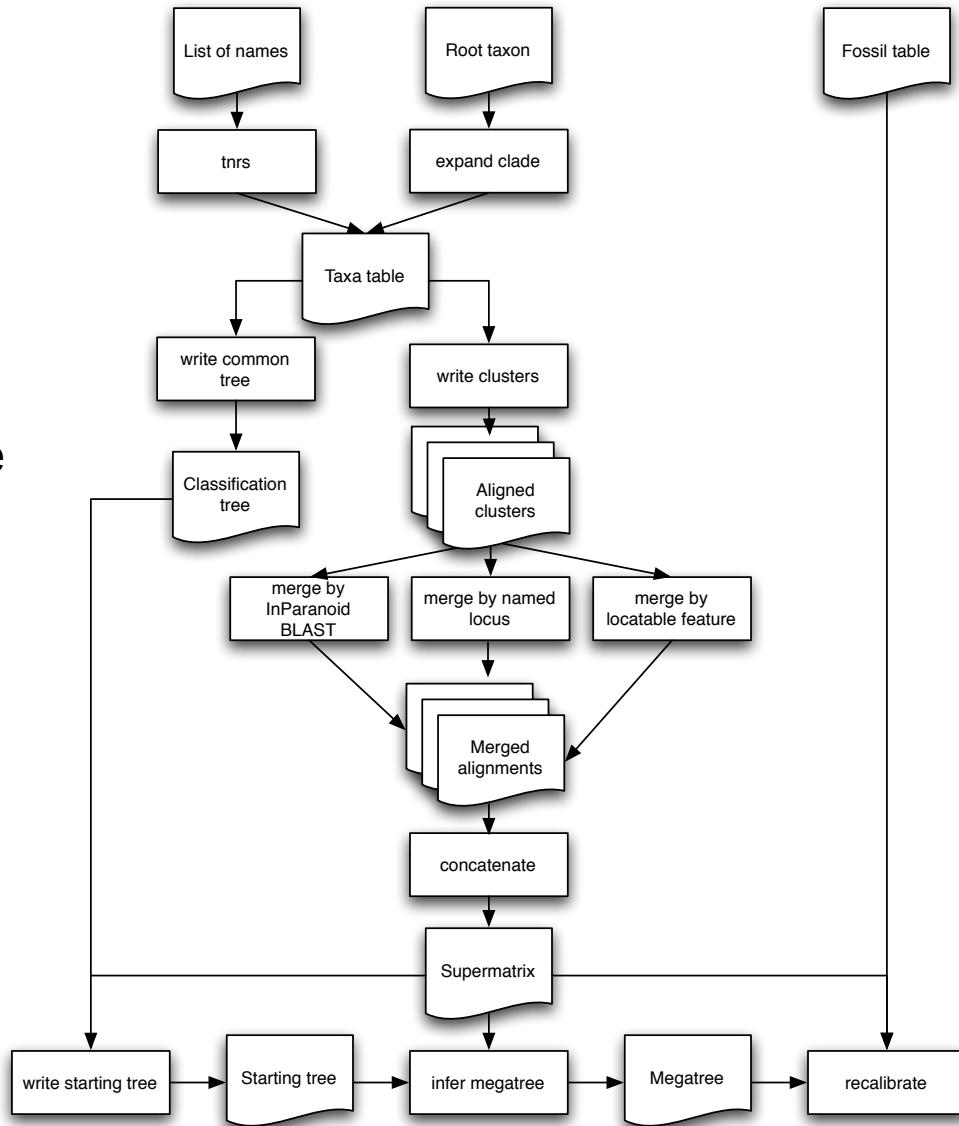
# Usage scenarios

- User wants a chronogram for an entire clade, e.g. Gentianales
- User wants a chronogram for a selected list of species, e.g. all those within a geographical area
- We want to assemble megatrees for integration downstream, e.g. to OpenTOL

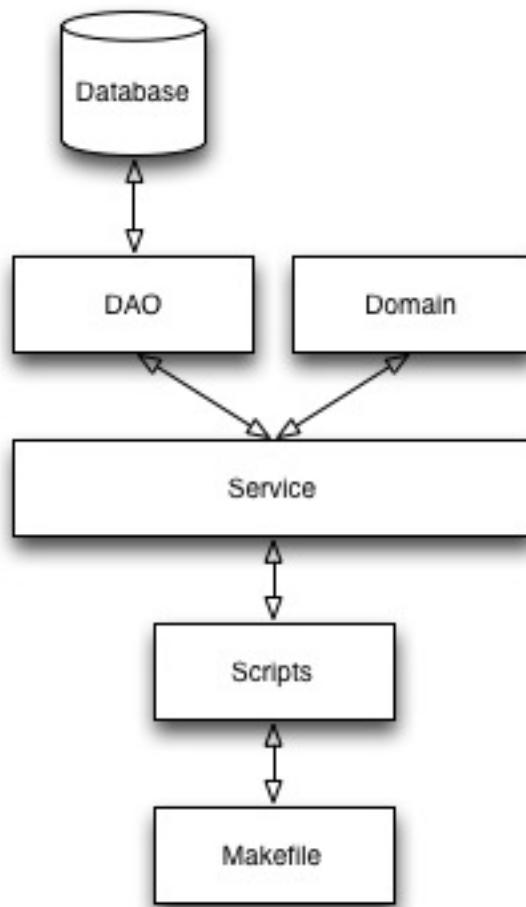


# Pipeline steps

1. create a table of NCBI taxon IDs of interest
2. write out and align the PhyLoTA clusters for these taxa
3. since PhyLoTA clusters are limited by taxonomic breadth, somehow merge orthologous clusters (first: figure out whether they are orthologous)
4. concatenate the merged clusters
5. generate a starting tree
6. run tree search

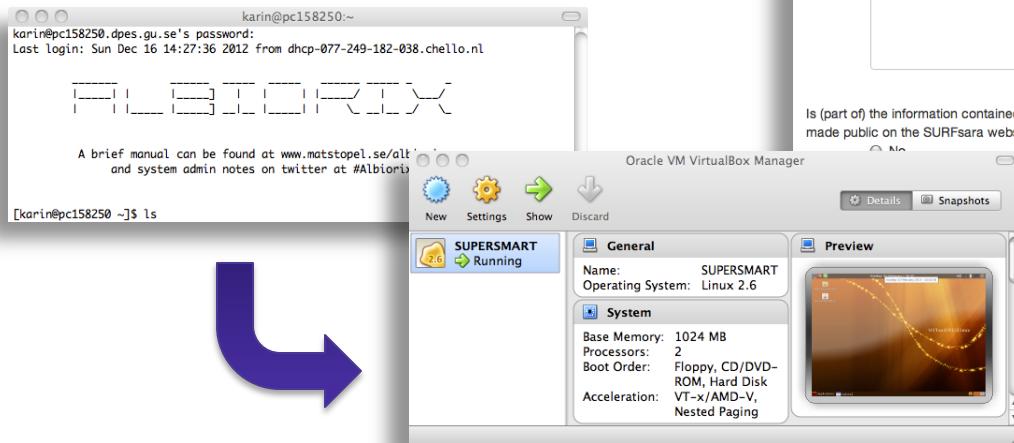


# RDBMS architecture



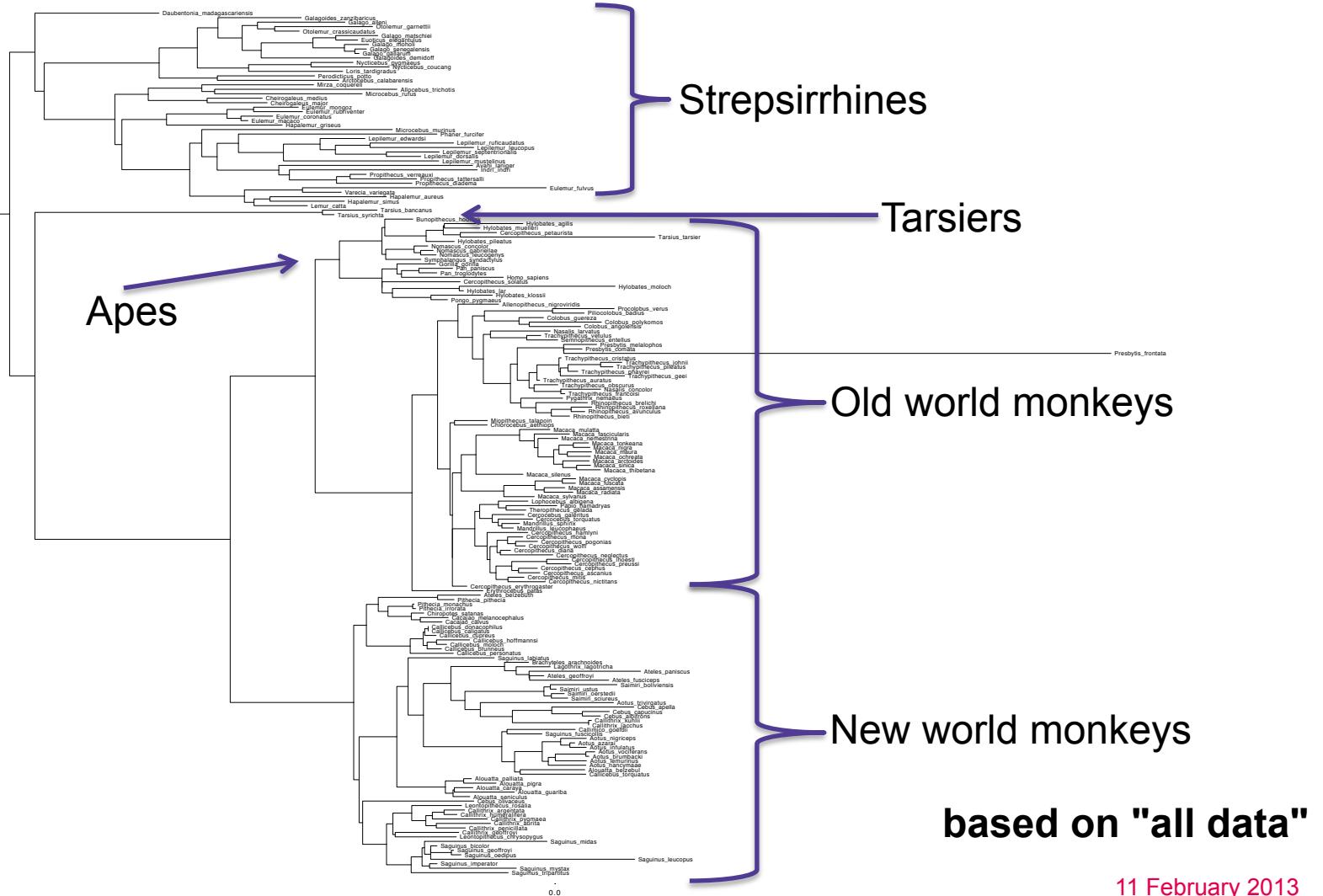
# Infrastructural needs

- A unix-like environment, tested on MacOSX, various Linux flavours
- MPI architecture (openMPI)
- MySQL database
- Standalone BLAST, exaML, phytime, muscle
- Perl with dependencies

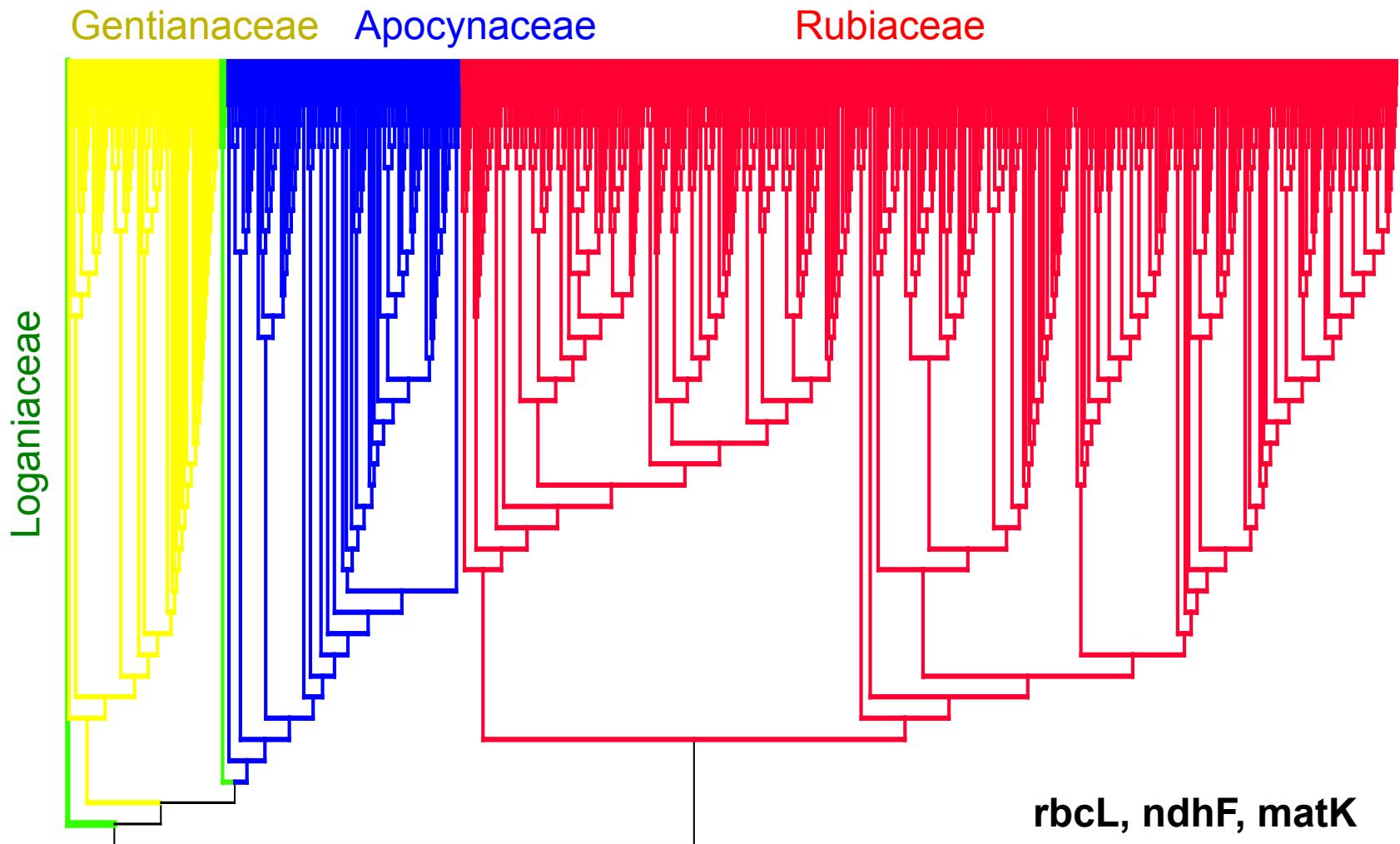


A screenshot of a web browser displaying the 'SURF SARA Resource request form' at <https://www.surfsara.nl/e-infra/>. The page has tabs for Personal, Resources, and Project, with 'Resources' selected. It contains fields for 'Title of your project', 'Describe your request', and 'Please describe the anticipated results of the request'. A question at the bottom asks if the information is confidential, with 'No' checked.

# **Examples of *preliminary* results: Primates**



# Examples of *preliminary* results: Gentianales



# To do / future work

- Support for more types of sequences (so far: CDS, misc\_RNA). For example: rRNA
- Bootstrapping, i.e. write out bootstrapped versions of supermatrix, run independent searches on these, compute consensus
- Recalibration based on fossils (e.g. treePL, r8s, BEAST)



# Points of discussion

## Design

- Locus selection
- Taxonomic reconciliation
- Orthology assignment
- Divide and conquer strategies
- Tool choice (aligners, tree-searchers, branch length scalers)
- Feature wish-list
- Other use-cases

## Sustainability

- Infrastructure
- Funding
- Developers
- Collaborations (BioVeL, OpenTOL, PhyloTastic)

## Outreach

- Website
- Publications
- Workshops
- Courses





# Thanks!

Naturalis  
Biodiversity  
Center

