A rough guide to SNAPP

Remco Bouckaert, David Bryant
remco@cs.auckland.ac.nz,david.bryant@otago.ac.nz
Auckland University

March 1, 2016

1 Introduction

SNAPP is a program to perform MCMC analysis on SNP and AFLP data using the method described in [2]. It is released under the lesser GPL licence. Source code is available on request to the authors. Executables for Mac, Linux and Windows are available for download from `http://snapp.otago.ac.nz/`.

2 Typical usage

In summary:

- Prepare input file, setting data, priors, initial tree and mcmc operators and options. See Section 4 for details on settings and Section 4.1 for details on how to use Beauti to set up an analysis file for SNAPP.
- Run SNAPP (see Section 5 for details).
- Analyze the trace, e.g. using `Tracer`[1] (see Section 6 for details), and depending on whether the chain converged, rerun with longer chain or other parameters.
- Analyze tree file by using the TreeSetAnalyser and DensiTree programs packaged with SNAPP, or alternatively using `splitstree`[2], `figtree`[3] (see Section 7 for details).

[1] Part of BEAST, or separately available from `http://beast.bio.ed.ac.uk/Tracer`
[2] Available from `http://www.splitstree.org/`
[3] Available from `http://tree.bio.ed.ac.uk/software/figtree/`

3 Preliminaries: variables and parameters

SNAPP is an MCMC program that takes AFLP (or SNP) data and produces a sample from the posterior distribution of species trees (and their parameters). There are four kinds of variable in the sample:

- The species tree topology;
- The branch lengths in the species tree, or equivalently, the divergence times for nodes in the species tree;
- The forward and backward mutation rates;
- The effective population size for each ancestral population in the species tree, given in terms of θ.

A real source of confusion, both for SNAPP and similar methods, is that the rates of mutation and times are confounded, so we typically rescale the units of time. To make this explicit, here are some of the standard population parameters and their units:

μ	mutation rate	expected number of mutations per site per generation
g	generation time	length of generation time, in years
N	effective population size	number of individuals
θ	average divergence	expected number of mutations between two individuals

For a diploid population, $\theta = 4N\mu$. If μ is instead the expected number of mutations per site per year then we would have $\theta = 4N\mu g$.

In the analysis in [2] we measured time in terms of expected number of mutations. Hence

- The expected number of mutations per unit time is 1;
- A branch of length τ in the species tree corresponds to τ/μ generations, or $\tau g/\mu$ years.
- The backward and forward mutation rates, u and v, were constrained so that the total expected number of mutations per unit time is 1, which gives
$$\frac{v}{u+v}u + \frac{u}{u+v}v = 1.$$
- The θ values are unaffected by this rescaling. If the true mutation rate μ is known, then the θ values returned by the program can be converted into effective population sizes using
$$N = \theta/(4\mu).$$

4 Preparing input file

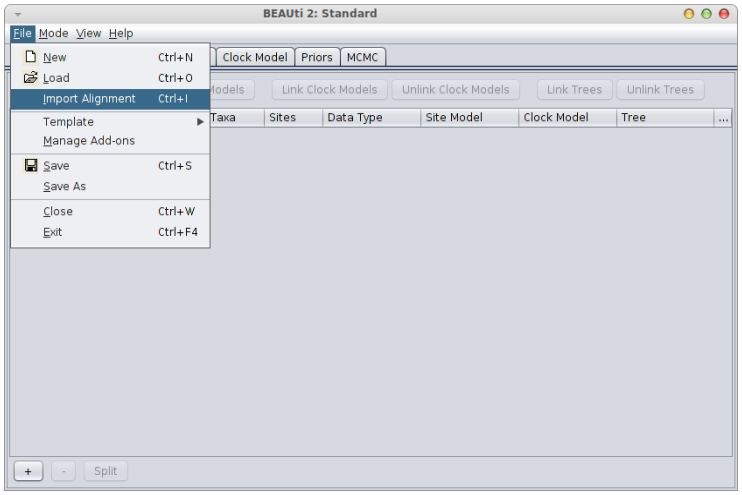
SNAPP uses an XML file as its specification. The easiest ways to create this XML file are to (i) use Beauti to create an XML file as outlined in Section 4.1, or (ii) export an appropriately formatted snaphyl file from SplitsTree.

To prepare the input file, the following items need to be specified: SNP or AFLP data, prior parameters, starting tree and MCMC parameters, as explained below.

4.1 Using Beauti

After starting Beauti, the first thing to do to specify an analysis is selecting the SNAPP template by selecting the menu File/Templates/SNAPP. Next thing to do is importing an alignment of SNP or AFLP data. Select the File/Import alignment menu item and a file chooser pops up with which the file containing an alignment in Nexus format can be selected. If an analysis file was saved before, it can be loaded back into Beauti by using the File/Load menu.

Figure 1: Importing an alignment in Beauti



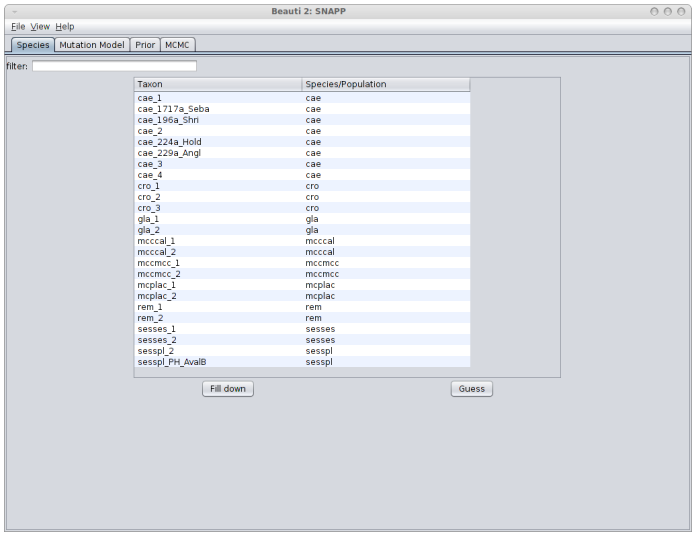
4.1.1 Beauti taxon set editing

When loading an alignment, Beauti guesses which of the alignments form a species by looking at the separator characters ('.','-','_' and ',') in taxon names. However, you should carefully go through the taxon sets and make corrections if necessary. Taxon sets can be renamed by editing the labels in the tree. Taxa can be moved around using drag an drop. To highlight taxa names matching part of a name, the filter entry at the top of the tree can be used. All taxon names matching[4] the text in the filter are highlighted, the remainder greyed

[4] A regular expression of the form ".*[filter entry text].*" is used for matching.

out.

Figure 2: Editing taxon sets in Beauti



4.1.2 Beauti mutation model settings

Click the tab 'mutation model' to show u, v and coalescent rate parameters and a number of optional check boxes. By default, the MCMC does not sample values for the backwards and forward mutation rates, u and v. These are fixed at their initial values. We note that, given these rates, the stationary frequencies of the two states are $\pi_0 = v/(u+v)$ and $\pi_1 = u/(u+v)$, while the mutation rate is

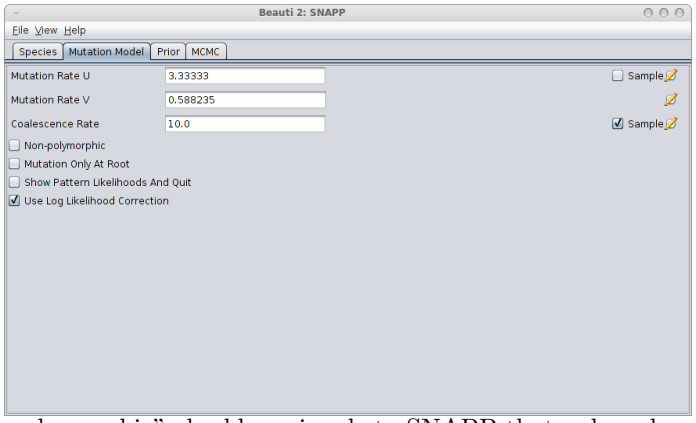
$$\pi_0 u + \pi_1 v = \frac{2uv}{u+v}.$$

Hence, given an estimate for π_0, and the constraint that the mutation rate is 1, we can solve for u and v as

$$\begin{aligned} u &= \frac{1}{2\pi_0} \\ v &= \frac{1}{2\pi_1} \\ &= \frac{1}{2(1-\pi_0)}. \end{aligned}$$

A fixed value can be entered, which will be used as initial value. When clicking the 'sample' check box, the parameter will be estimated during the course of the MCMC chain. Click the little button with the 'e' on the right to specify bounds on a parameter that is sampled. When sampling a parameter a prior on the parameter need to be specified (which will show up automatically under the 'prior' tab).

Figure 3: Mutation model parameters in Beauti



The "non-polymorphic" checkbox signals to SNAPP that only polymorphic (variable) sites have been included in the data (all constant sites are ignored). It is important to check this box if SNP data is being used, as the likelihood calculations are quite different if SNAPP assumes all constant sites have already been removed.

The "mutationOnlyAtRoot" checkbox indicates conditioning on zero mutations, except at root (default false). As a result, all gene trees will coalesce in the root only, and never in any of the branches.

It is possible to indicate whether alleles are dominant, however in our experience this only leads to much longer run times without changing the analysis significantly. Therefore it is set to false by default.

4.1.3 Beauti tree prior settings

Click the 'prior' tab to show the prior used in the analysis. SNAPP uses a *Yule prior* for the species tree and branch lengths on the species tree. This prior has a single parameter, λ, which governs the rate that species diverge. This rate, in turn, determines the (prior) expected height of the species tree, which we denote by r. The formula connecting these is two quantities is

$$\lambda = \frac{1}{r}\left(\sum_{k=1}^{n-1} \frac{k}{n(n-k)}\right)$$

where n is the number of species, which can be used to specify the λ value giving the desired root height.

4.1.4 Beauti rate prior settings

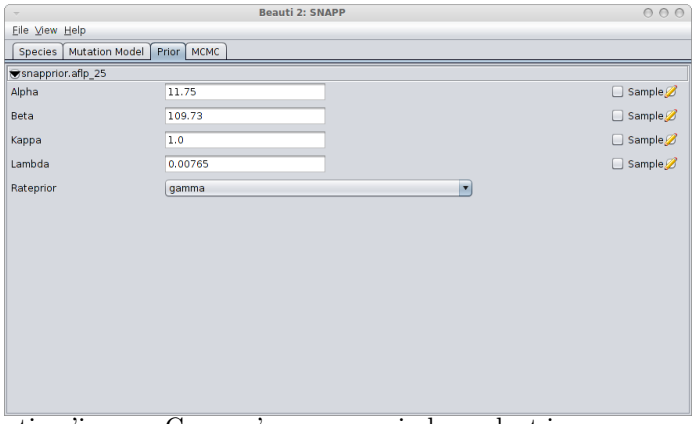
The theta (θ) values for the ancestral population sizes all have gamma distribution priors by default. This can be altered by selecting the combobox and selecting one of 'gamma', 'inverse gamma', 'CIR' or 'uniform'. The parameters

for the gamma distributions are the same over the entire tree, and for all trees, and the prior distribution for each ancestral population are independent. There are two parameters determining the gamma distribution:

- the shape parameter α;
- the scale parameter β.

Note that the mean of the prior distribution if α/β while the variance of the prior distribution is $\alpha/(\beta)^2$.

Figure 4: Prior settings in Beauti



When selecting 'inverse Gamma' we assume independent inverse gamma distributions for thetas, so $(2/r)$ has an inverse gamma (alpha,beta) distribution. That means that r has density proportional to $1/(r^2)*INVGAMMA(2/r|alpha, beta)$.

When selecting the CIR process [6], the kappa parameter is used as well. The CIR process has SDE

$$dr = \kappa(\theta - r)dt + \sigma\sqrt{r}dz_1$$

has a stationary distribution that is gamma with

$$\alpha = 2\kappa\theta/\sigma^2$$

and

$$\beta = 2\kappa/\sigma^2$$

The correlation between time 0 and time t is $exp(-\kappa t)$.

Let $c = 2\kappa/(\sigma^2 * (1 - exp(-\kappa t)))$

If we condition on rate $r0$ at time 0, the distribution of $2*c*rt$ is non-central chi squared with

$$2q + 2 = 4\kappa\theta/\sigma^2$$

degrees of freedom and parameter of non-centrality

$$2u = cr_0exp(-\kappa t)$$

Converting these into our set of parameters (alpha, beta, kappa) we have

$$\theta = \alpha/\beta$$

$$\sigma^2 = 2\kappa/\beta$$

so

$$c = \beta/(1 - exp(-\kappa t))$$

$$df = 2q + 2 = 2\alpha$$

$$nc = 2u = 2\beta r_0 \frac{exp(-\kappa t)}{1 - \exp(-\kappa t)}$$

We are applying the CIR process to

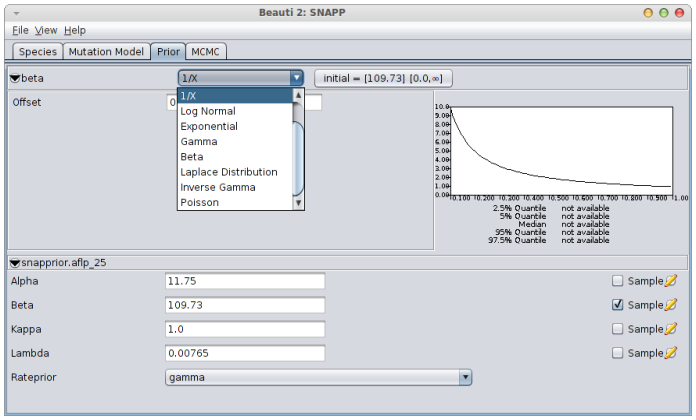
$$\theta = 2/r$$

When selecting 'uniform' we assume the rate is uniformly distributed in the range 0...10000, which means there is a large proportion of the prior indicates a large value, with a mean of 5000.

4.1.5 Beauti hyper prior settings

When sampling any of the parameters of the prior, hyper priors need to be specified. These show up automtically when clicking the 'sample' checkbox on a parameter. In the following figure, alpha and beta and selected for sampling. Various hyper prior distributions can be specified for these parameters.

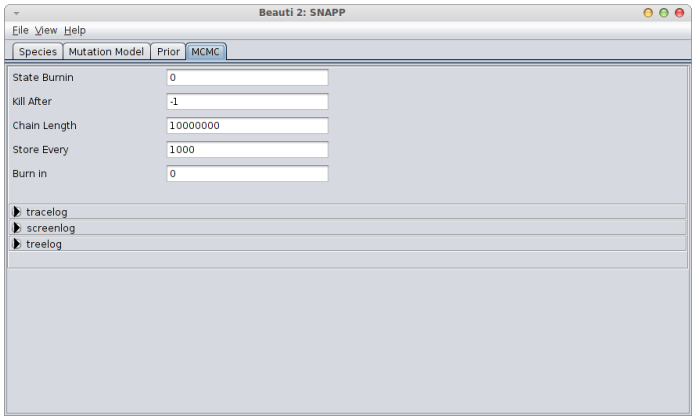
Figure 5: Hyper-prior settings in Beauti



4.1.6 Beauti MCMC settings

Click the 'MCMC' tab to select parameters for the MCMC algorithm. Under the log entries, file names and logging frequency can be specified

Figure 6: MCMC settings in Beauti

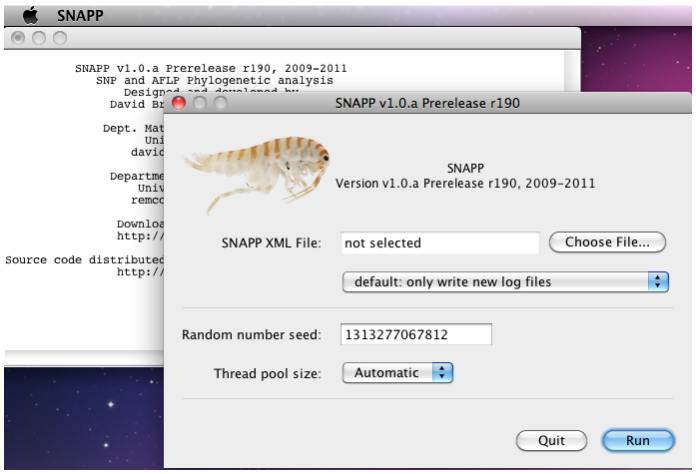


5 Launching SNAPP

SNAPPcan be launched from the command line, or as a GUI application by double clicking the appropriate application.

When launching SNAPPas GUI app, a window and dialog pop up as shown in Figure 7.

Figure 7: Launching SNAPP



5.1 Command line options

```
Usage: snapp [-window] [-options] [-working] [-seed] [-prefix <PREFIX>] [-overwrite] [-res
  -window Provide a console window
  -options Display an options dialog
  -working Change working directory to input file's directory
  -seed Specify a random number generator seed
  -prefix Specify a prefix for all output log filenames
  -overwrite Allow overwriting of log files
  -resume Allow appending of log files
  -errors Specify maximum number of numerical errors before stopping
  -threads The number of computational threads to use (default auto)
  -help Print this information and stop

Example: snapp test.xml
Example: snapp -window test.xml
Example: snapp -help
```

Some more detail on the options:

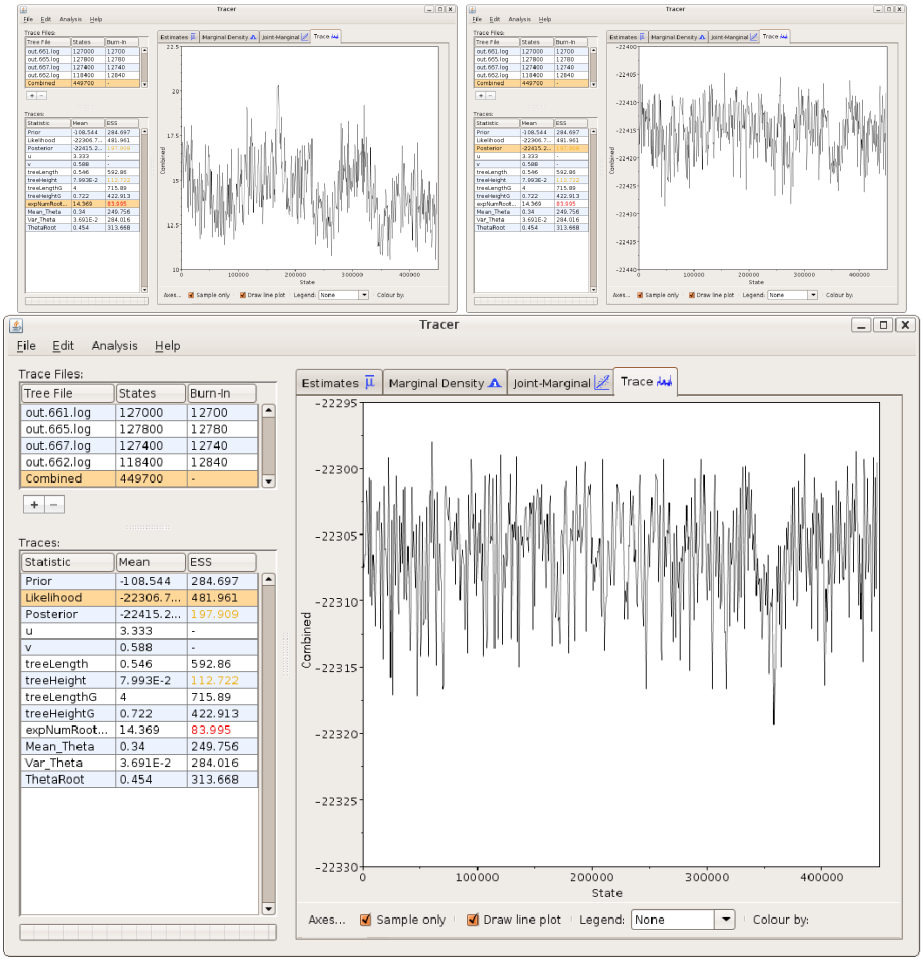
`-seed` is the random number seed, which defaults to 667. Repeated runs with the same seed result in the same outcomes. Running multiple chains with different seeds is recommended to test for convergence.

`-threads` specifies number of threads used. Increasing the number of threads will help on a multi-core machine. Though the performance does not increase linearly with the number of thread, it does improves considerably, especially on many-core CPUs like the Intel i7.

6 Analyzing the Trace

Trace files generated by SNAPP can be inspected using Tracer, which is part of BEAST. Figure 8 shows various screen shots of Tracer. The left hand (bottom) side shows mean and effective sample size (ESS) for various variables. An ESS of less than 100 is generally an indication that the chain has not converged for that variable yet (expNumRootLineages in Figure 8). When ESS is larger than 100, the next thing to check is the shape of the trace. If there are any obvious trends visible, or large fluctuations, this is generally a sign that the chain has not converged. For example the middle plate of Figure 8 shows the posterior trace showing signs that it jumps between two states. However, if the chain converged, it should look closer to something like a hairy caterpillar shown at the bottom of Figure 8.

Figure 8: Various traces, from unacceptable to ok.



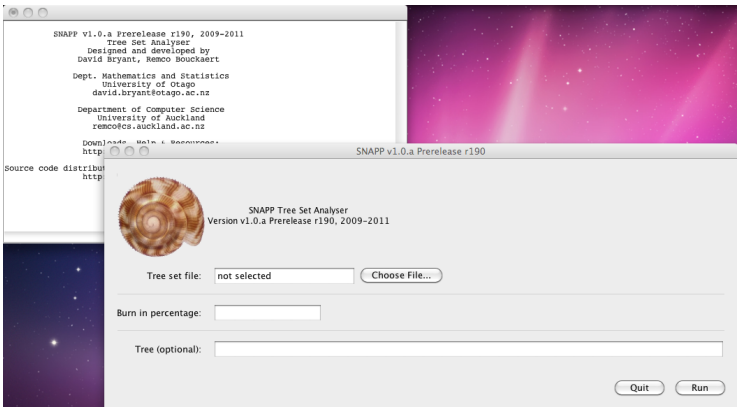
7 Analyzing the Trees

SNAPP can generate a tree set containing species trees. There are various ways to analysing these tree sets.

- calculate a summary tree, e.g. through the tree annotator in BEAST. The benefit is that a single tree is produced with 'error bars' on the position of internal nodes. However, when there is a lot of uncertainty in the topology of the species tree, this may not be directly obvious from the summary tree and some skill is required to recognise these kind of situations.
- calculate a set of clades with high frequency, e.g. through the tree log analyser in BEAST. This will show a set of clades without the uncertainty in node heights and the set can become quite large and hard to interpret.
- Draw a consensus tree or consensus network (as in SplitsTree [3]), which is a graph containing edges wherever such edges appear (possibly at some threshold frequency) in the tree set.
- Use multidimensional scaling (MDS) as for example implemented in [5]. MDS allows identification of tree islands in a compelling way, but uncertainty of node heights is hard to interpret.
- Draw a DensiTree [4], which shows the complete tree set drawn transparently in a single image. Areas with hight uncertainty in the tree topology will be easy to spot and distribtuion in node heights shows up clearly as well.

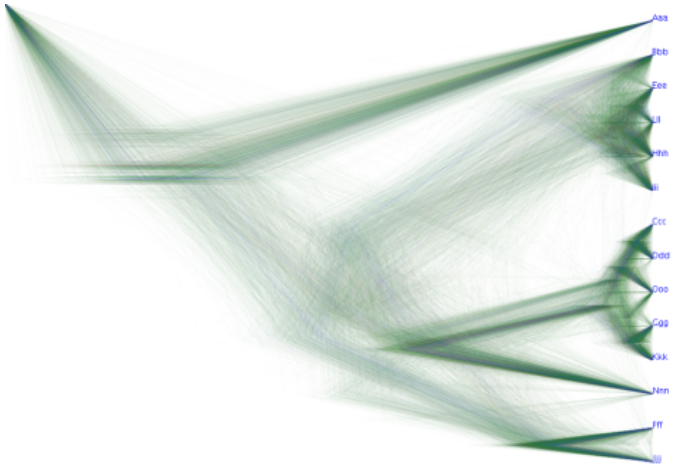
`TreeSetAnalyser` is a program that is part of the SNAPP distribution. It reads in a tree set file produced by SNAPP and determines the distribution of topologies of the species tree. Furthermore, it determines whether a pre-defined tree topology (if any is specified) is in the 95% HPD set of topologies. To start `TreeSetAnalyser`, double click the TreeSetAnalyser icon and the following window should pop up.

Figure 9: Tree set analyser



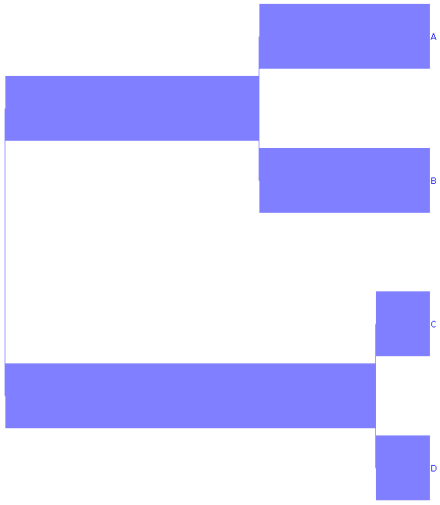
By default, the burn in percentage is 10%. The Tree entry allows to specify a tree topology in Newick format. When clicking Run, the results of the analysis are shows in the text window.
Figure 10 shows an example of a DensiTree.

Figure 10: Example of a DensiTree. This highlights the uncertainty in topology of the two 5-taxon clades. Also, note the increasing uncertainty in internal node height going from the taxa to the root.



One feature of `DensiTreeS` is that it allows the branch thicknesses in a tree to represent a parameter associated with a branch, such as θ. This figure can be exported as SVG file and manipulated in a drawing program (e.g. PhotoShop, Gimp) to annotate a tree with extra information to produce high quality ready for print images.

Figure 11: Example of a DensiTree showing only the dominating consensus tree where branch thickness represents θ associated with the branch.



References

[1] A Drummond and A Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology, Jan 2007.

[2] David Bryant, Remco Bouckaert, Joseph Felsenstein, Noah Rosenberg, Arindam RoyChoudhury. Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis. Mol. Biol. Evol. 29(8):1917-1932, 2012

[3] D. H. Huson and D. Bryant. Application of Phylogenetic Networks in Evolutionary Studies, Mol. Biol. Evol., 23(2):254-267, 2006. `http://www.splitstree.org/`

[4] Remco R. Bouckaert. DensiTree: making sense of sets of phylogenetic trees Bioinformatics, Vol. 26, No. 10. (15 May 2010), pp. 1372-1373.

[5] Hillis, D.M., Heath, T.A. and St John, K. Analysis and Visualization of Tree Space. Syst. Biol. 54(3):471-82, 2005.

[6] Cox, J.C., J.E. Ingersoll and S.A. Ross. "A Theory of the Term Structure of Interest Rates". Econometrica 53: 385407, 1985.